IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 25, 2021, accepted April 7, 2021, date of publication April 13, 2021, date of current version May 24, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3072900

Efficient Automated Processing of the Unstructured Documents Using Artificial Intelligence: A Systematic Literature Review and Future Directions

DIPALI BAVISKAR[1], SWATI AHIRRAO[1], VIDYASAGAR POTDAR[2], AND KETAN KOTECHA[3]

[1]Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune 412115, India
[2]Blockchain Research and Development Laboratory, Curtin University, Perth, WA 6845, Australia
[3]Symbiosis Centre for Applied Artificial Intelligence, Symbiosis International (Deemed University), Pune 412115, India

Corresponding author: Swati Ahirrao (swatia@sitpune.edu.in)

This work was supported by the Research Support Fund of Symbiosis International (Deemed University).

ABSTRACT The unstructured data impacts 95% of the organizations and costs them millions of dollars annually. If managed well, it can significantly improve business productivity. The traditional information extraction techniques are limited in their functionality, but AI-based techniques can provide a better solution. A thorough investigation of AI-based techniques for automatic information extraction from unstructured documents is missing in the literature. The purpose of this Systematic Literature Review (SLR) is to recognize, and analyze research on the techniques used for automatic information extraction from unstructured documents and to provide directions for future research. The SLR guidelines proposed by Kitchenham and Charters were adhered to conduct a literature search on various databases between 2010 and 2020. We found that: 1. The existing information extraction techniques are template-based or rule-based, 2. The existing methods lack the capability to tackle complex document layouts in real-time situations such as invoices and purchase orders, 3. The datasets available publicly are task-specific and of low quality. Hence, there is a need to develop a new dataset that reflects real-world problems. Our SLR discovered that AI-based approaches have a strong potential to extract useful information from unstructured documents automatically. However, they face certain challenges in processing multiple layouts of the unstructured documents. Our SLR brings out conceptualization of a framework for construction of high-quality unstructured documents dataset with strong data validation techniques for automated information extraction. Our SLR also reveals a need for a close association between the businesses and researchers to handle various challenges of the unstructured data analysis.

INDEX TERMS Artificial Intelligence (AI), document analysis, information extraction, named entity recognition (NER), optical character recognition (OCR), robotics process automation (RPA), unstructured data.

I. INTRODUCTION

With the advent of new communication media and various applications like social media, mobile applications, and digital marketing, the data produced does not have a typical format or predefined schema like the standard data and cannot be managed with the relational database models. Data is generated in various forms such as text, audio, videos, emails, and images. These are examples of unstructured data. Such data lacks structure and is not standardized [1], which makes editing, searching, and analysis difficult for the majority of the organizations [2].

Forbes statistics [3] states that analyzing unstructured data is an issue for 95% of business organizations, as they do not have the required expertise to deal with the unstructured data. Over 150 trillion gigabytes (150 zettabytes) of unstructured data would need to be analyzed by 2025. The organizations can use data analysis tools to better understand the customer needs and forecast market variations. In simple words, there are countless applications of unstructured data. By 2022, the yearly profits of the ‘‘global unstructured data and business analytics’’ market are estimated to be 274.3 billion U.S. dollars [4].

The associate editor coordinating the review of this manuscript and approving it for publication was Shadi Alawneh.

TABLE 1. Various application domains for automatic information extraction from unstructured documents.

Application Areas	Usage
Healthcare	Patient details and disease extraction from Electronic Health Record (EHRs).
HR	Employee details extraction and verification from their documents, Payroll process automation, Job recommendations or hiring shortlisted candidates from the collection of CVs.
Insurance	Automating Claims Processing & clearance form filling, Extraction of policy premium amount details from policy documents.
Travel Sector	Extracting and verifying ticket booking and traveler details.
Banking and Financial Services	KYC Verification, Auto form-filling for cards activation, Fraud claim detection, Pattern discovery from the customer documents.
Government	Automating address change requests, License renewal by verifying the user details.
Infrastructure	Infrastructure related document processing and information extraction from documents, Invoice and receipt digitization.
Legal	Contract element extraction from legal documents, identifying clauses, and involved risks.

Nearly 80% of the generated data in an organization is unstructured [2]. This data is essential to the organization in decision-making, predictive analysis, and pattern-finding tasks [1]. Effective processing of such unstructured data is always challenging, time-consuming, and expensive for any organization [2]. As the data is generated at an exceptional speed, the valuable information hidden in them cannot be made useful, unless there is some form of automated analysis. Table 1. shows various applications domains for automatic information extraction from the unstructured documents [1], [5], [6].

Automation helps organizations to organize and access useful information in a structured manner [6]. Automation of the unstructured data stored in the digital format would allow the organizations to quickly gain insight into their businesses, increase their competitive edge, improve their productivity, and make innovations. The organizations thus adapt to the automation solutions by understanding the importance of Artificial Intelligence-based (AI-based) technologies such as Computer Vision (CV) and Natural Language Processing (NLP). AI technologies can understand and classify unstructured data like text, images, and scanned documents, better than traditional information extraction methods [5], [6].

Increasing volume and the need for effective use of the unstructured data necessitate developing an AI-based unstructured document processing framework, that would help the organizations automatically get the insights from unstructured data. Thus, this is considered as a significant and upcoming research area.

A. SIGNIFICANCE AND RELEVANCE

Unstructured data is an integral part of many organizations. The forms such as invoices, customer details, insurance claims serve as proof and records of transactions and other critical events in the organization. Proper processing of these forms is essential. Manual processing is likely to slow down the process and may result in errors and delays. Automatic data extraction tackles these issues, giving an automated and digitized solution to document processing. Automating the time-consuming and repetitive tasks helps to improve the productivity and growth of the organization. AI enables efficient and automatic extraction of useful information from unstructured documents. AI also helps to create a more understandable analysis of the unstructured documents that the organizations may use in their critical decision-making process.

B. EVOLUTION OF THE TECHNIQUES USED FOR AUTOMATIC INFORMATION EXTRACTION FROM UNSTRUCTURED DOCUMENTS

Figure 1. shows the evolution of the techniques used for automatic information extraction from unstructured documents. Earlier, the organizations employed a manual workforce to do data entry, process paper-based documents, and supply the needed information to the next business processing chain.

As the first step towards digitization, the organizations started using Optical Character Recognition (OCR). OCR was utilized to transform the scanned document contents into digital format. The preliminary versions of OCR are required to provide each character image and limited to recognize only one font at one time. In the early 2000s, ‘‘Omni-font OCR’’ was proposed, which could process text printed in almost any font [7]. Later, it became a cloud-based service, which could be accessed via desktop and mobile applications. Today, many OCR service providers offer OCR technology via APIs and are proficient in recognizing almost all the characters and fonts to a reasonable accuracy level.

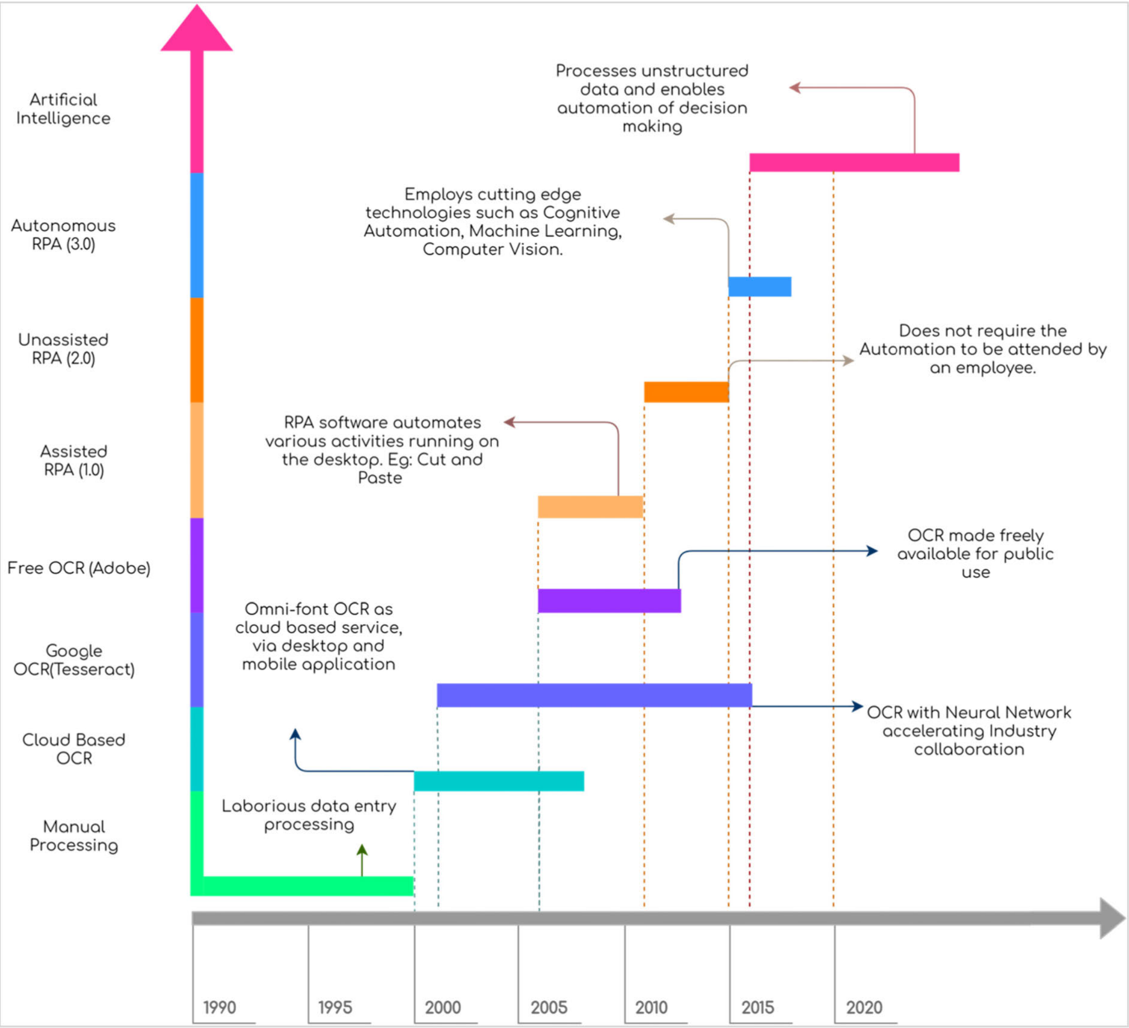

Artificial Intelligence
Autonomous RPA (3.0)
Unassisted RPA (2.0)
Assisted RPA (1.0)
Free OCR (Adobe)
Google OCR(Tesseract)
Cloud Based OCR
Manual Processing
Processes unstructured data and enables automation of decision making
Employs cutting edge technologies such as Cognitive Automation, Machine Learning, Computer Vision.
Does not require the Automation to be attended by an employee.
RPA software automates various activities running on the desktop. Eg: Cut and Paste
OCR made freely available for public use
Omni-font OCR as cloud based service, via desktop and mobile application
OCR with Neural Network accelerating Industry collaboration
Laborious data entry processing
1990
1995
2000
2005
2010
2015
2020

FIGURE 1. Evolution of the techniques used for information extraction from unstructured documents.

Moving further in automation, organizations started using Robotics Process Automation (RPA) to replace rule-based, structured, and repetitive processes with a software bot [8]. Rule-based logic is at the core of the most automated processes. It is a logical program based on predefined rules to perform automated actions. RPA has evolved from RPA 1.0, RPA 2.0, and RPA 3.0.

Assisted RPA or RPA 1.0 automates several user actions and applications, that are being executed on the user's computer. Automating a simple job like a cut, copy, and paste of information from one computer system to the other computer system is an example of RPA 1.0. However, RPA 1.0 must be applied, when human-computer communication is essential [2], [9].

Unassisted RPA or RPA 2.0 can be installed on several machines to automate the task without human assistance. It can significantly reduce human interaction with the business processes [9]. An employee logging into a system, activates the initiation of the processes, notices their execution, and shuts down the system, when it is finished. This process can be automated entirely using RPA 2.0 without any human intervention.

Autonomous RPA or RPA 3.0 is the latest version of RPA. It takes the advantage of AI and CV. As technology progresses, RPA had also seen more beneficial improvements [9]. An example of RPA 3.0 could be segregation and automatic response to several emails. This AI-enabled RPA is sometimes referred to as Cognitive-RPA [10].

Nowadays, AI-based automation uses several promising approaches like NLP, Machine Learning (ML), and Text Analytics, formulating the equal usefulness of structured as well as unstructured data. With AI, unstructured data can be analyzed, managed, and processed to get valuable insights with less human efforts and interventions [2].

II. PRIOR RESEARCH

One of our SLR objectives is to develop a clear and detailed understanding of the existing automatic information extraction techniques for unstructured documents. As far as we know, there are very few Systematic Literature Reviews (SLR) available in this research area.

The study [11] is one of the recent and significant SLR providing a good overview of text analytics for unstructured data processing in the financial domain. The study added value to the literature, by offering valuable insights from unstructured document processing for the finance industry. The authors discussed how text analytics could help customer onboarding, predict market variations, fraud detection and prevention, improve operational activities, and develop innovative business models. The authors highlighted two important unstructured data sources for the text analytics used in the financial sectors: outside data sources and inside data sources. Inside data sources includes log files data, transaction data, and application data. Outside data sources includes any social media data, and website data. The survey also provides the useful text analytics methods such as sentiment analysis, Named Entity Recognition (NER), topic extraction, and keyword extraction. However, the survey lacks a detailed discussion of the existing information extraction techniques for automating data extraction from the unstructured documents.

The surveys [1], [6] cover the challenges of different types of unstructured data like images, text, audio, and video. The authors highlighted the unstructured data challenges such as the representation and conversion of the unstructured data into structured data, massive growth in its volume, and heterogeneous data types.

The survey [1] presents the information extraction techniques for the unstructured data. The authors concluded that Deep Learning has generalizability and adaptability features. So it could be applied with the traditional information extraction techniques to manage the unstructured data well. However, the survey does not provide details of any tool or framework used for the information extraction. The survey lacks a structured approach and methods for the specific unstructured data type. The authors reported in their future work, that researchers may focus on the data pre-processing techniques for improvement in data quality. Our SLR discusses few data pre-processing techniques, which will enhance the performance of the model. The survey [6] presents an overview of the information extraction techniques for various types of unstructured data like images, text, audio, and video. The authors also investigated the limitations of the existing information extraction techniques due to the variation and size of the unstructured data. However, the discussion on any information extraction model or framework to improve the existing information extraction methods is not covered. Our SLR is more focused, and niche as ''unstructured text'' data is specifically attended. NER is a form of NLP and a sub-field of AI used in the information extraction tasks. A named entity is a real-world entity, such as date, name, organization, location, and products that can be denoted with a proper noun. The survey [12] highlights recent advances in Named Entity Recognition (NER), Named Entity Disambiguation (NED) and Named Entity Linking (NEL) using Knowledge Graphs (KG). Here, named entities are designated as the nodes in a graph, and edges represent the semantic relationship of the nodes. However, the author has not concentrated on NER from the unstructured documents. The survey [13] focuses on the clinical information extraction applications, using Electronic Health Records (EHR). The survey presents a few frameworks such as Unstructured Information Management Architecture (UIMA), General Architecture for Text Engineering (GATE), Medical Language Extraction and Encoding (MedLEE) for information extraction from EHR. The survey lacks a discussion on AI-based information extraction techniques.

The structured literature review [14] recognizes several current, RPA-related issues, themes, and challenges for future exploration. The survey focuses on, how RPA has seen significant acceptance in organizations, aiming to increase operational productivity. The authors highlighted the benefits of RPA, related to organizational performance improvement and cost reduction by reducing the human workforce in routine business processes and improving the work quality. The survey also reported various RPA vendors, who provide commercial RPA solutions. However, this survey lacks a discussion on AI-based approaches used in RPA for handling the unstructured data. Our SLR highlights the role of RPA and process selection criterion for RPA implementation.

We observe few limitations of the prior research work, which can be stated as follows:

1. Existing SLR are domain-specific or task-specific.
2. Existing literature does not examine the generalizability of the information extraction techniques to handle multiple layouts or formats of the unstructured documents. For example, each company or contractor possibly has its unique or custom format for invoices, and purchase orders. The documents that can be free-form and do not have a fixed structure are said to have multiple layouts.
3. Discussion on data validation techniques is not covered in the existing SLR.
4. Very few studies surveyed tools or frameworks available for the automatic extraction of information from unstructured documents.

Our SLR is exhaustive in terms of showcasing the current developments or trends and challenges related to the unstructured document processing, by attempting detailed investigation on AI-based information extraction techniques, availability and quality of publicly available datasets, data validation methods, and tools or frameworks used for information extraction. Our SLR presents a comparative analysis of OCR, RPA, and AI-based approaches for information extraction from unstructured documents. Our SLR highlights the future research directions by emphasizing the research gaps.

A. MOTIVATION

There is no existing SLR that focuses on the information extraction techniques covering their explicit benefits, limitations, taxonomies, and comparative analysis. Existing

TABLE 2. Research questions.

Number	Research Questions	Objective
RQ-1	What are the different challenges with information extraction techniques to deal with unstructured data?	Unstructured data does not have a predefined schema or format. It is, therefore, more challenging to analyze. The aim is to understand the challenges with the information extraction techniques to manage unstructured data.
RQ-2	What are the different datasets available for unstructured data processing?	Obtainability of a dataset comprising sufficient training and testing data is necessary for good research results. The aim is to explore the publicly available datasets for unstructured data, the application areas of datasets, their sources, the size of the dataset, and the description of the contents of the dataset.
RQ-3	What are different data validation techniques used for the quality assessment of data?	The aim is to study data validation techniques used for the quality assessment of data. The goal is to identify the best suitable data samples, that would contribute to the performance of the model.
RQ-4	What are the different AI approaches used for unstructured data processing?	To review the benefits, limitations, and comparative analysis of widely used AI-based information extraction techniques for unstructured data.
RQ-5	What are the different online tools available for automatic information extraction from unstructured documents?	To explore the existing tools or frameworks used (or developed) for information extraction from unstructured data and provide their comparative analysis.

literature lacks a comprehensive survey focused on the publicly available datasets, and data validation methods. The literature also lacks an exhaustive study on frameworks or tools for automatic information extraction from unstructured documents.

The essence of this SLR is to highlight the available facts regarding:

- The existing information extraction techniques and their limitations to process unstructured documents,
- Publicly available datasets for information extraction from unstructured documents,
- Data validation methods used for the quality assessment of the data,
- Tools or frameworks used for information extraction,
- The comparative analysis of OCR, RPA, and AI-based techniques.

Therefore, the proposed SLR aims to provide insights to researchers for developing efficient information extraction techniques for unstructured documents.

B. RESEARCH GOALS

Our SLR aims to identify and critically analyze the existing studies and their outcomes in the context of the formulated research question.

Table 2. shows the research questions that were prepared to make this SLR work more focused.

C. CONTRIBUTIONS OF THE STUDY

Following are the contributions of our Systematic Literature Review:

- We identified 83 primary studies on automatic information extraction from unstructured documents published from 2010 to 2020. Other researchers can use these studies to advance their work in this area.
- A comprehensive review of the availability and quality of publicly available datasets and data validation methods is performed.
- A suitable benchmark for comparative analysis of the widely used AI-based techniques used for automatic information extraction from unstructured documents is provided.
- A summary of the existing tools or frameworks available for automatic information extraction from unstructured documents is presented.
- The research gaps in this area were identified, which will help researchers and business organizations choose the proper method for automatically extracting valuable information from the unstructured documents using AI techniques. We discussed the future research directions in this area.
- A conceptualization of a framework for the construction of a high-quality unstructured documents dataset with strong data validation techniques for automated information extraction is provided as an outcome of SLR.

Figure 2. shows the outline of our SLR in different sections.

III. RESEARCH METHODOLOGY

SLR guidelines proposed by Kitchenham and Charters [15] were adhered for carrying out the detailed systematic literature review. Table 3. shows PIOC (Population, Intervention, Outcome, Context) approach published by Kitchenham and Charters [15] used for framing the research questions. Figure 3. presents the flowchart for the selection of relevant papers to answer our research questions.

A. SELECTION CRITERION FOR RESEARCH STUDIES

The keywords were chosen to get the desired search results, that would help to address the research questions. The

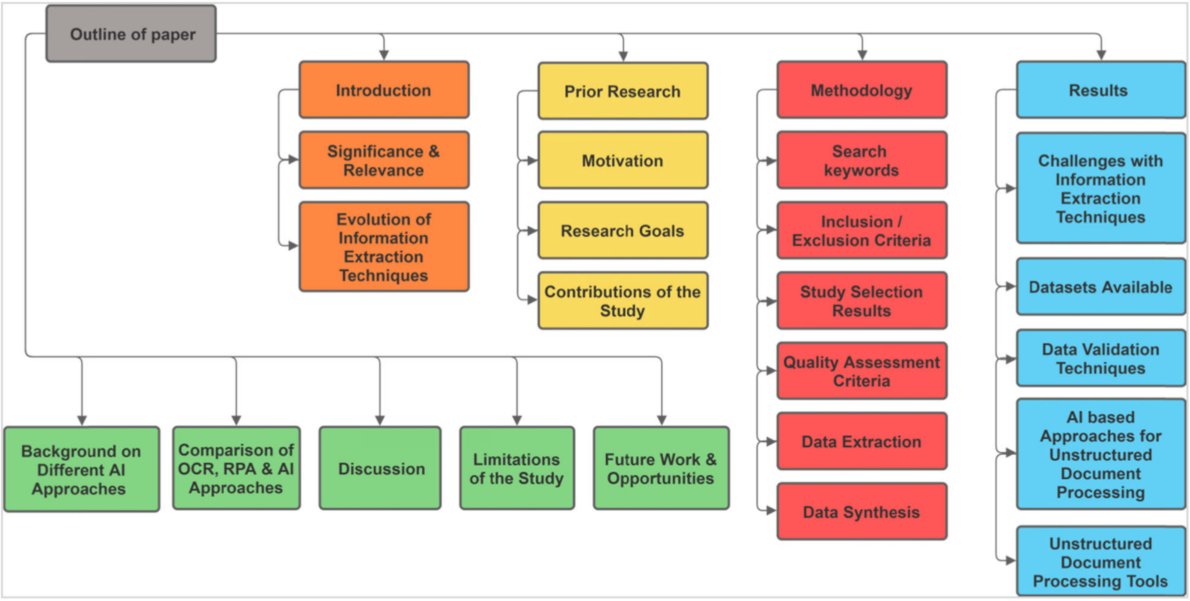

Outline of paper
Introduction
Significance & Relevance
Evolution of Information Extraction Techniques
Prior Research
Motivation
Research Goals
Contributions of the Study
Methodology
Search keywords
Inclusion / Exclusion Criteria
Study Selection Results
Quality Assessment Criteria
Data Extraction
Data Synthesis
Results
Challenges with Information Extraction Techniques
Datasets Available
Data Validation Techniques
AI based Approaches for Unstructured Document Processing
Unstructured Document Processing Tools
Background on Different AI Approaches
Comparison of OCR, RPA & AI Approaches
Discussion
Limitations of the Study
Future Work & Opportunities

FIGURE 2. Outline of paper.

TABLE 3. PIOC (Population, Intervention, Outcome, Context) criteria.

Parameter	Meaning	Keywords Used
Population	It is an application area	"document processing" OR "document analysis" OR "unstructured data" OR "big data"
Intervention	It is the software approach/tool or framework or model /technology/procedure that addresses a particular issue	"Artificial Intelligence" OR "AI" OR "Machine Learning" OR "Deep Learning"
Outcome	It must identify elements of significance to professionals, for example, high reliability, reduced manufacture costs, and reduced delay	"information extraction" OR "information retrieval" OR "named entity"
Context	It is the framework or context in which the intervention is provided	"Optical Character Recognition" OR "OCR" OR "Robotics Process Automation" OR "RPA"

following search string is used for searching the most relevant literature:

(“document* proces*” OR “document* analy*” OR “unstruct* data” OR “big data”) **AND**

(“Artificial Intelligence” OR “AI” OR “Machine Learning” OR “Deep Learning”) **AND** (“information extract*” OR “information retrieval” OR “nam* entit*”) **AND** (“Optical Character Recognition” OR “OCR” OR “Robotics Process Automation” OR “RPA”)

Table 4. shows the database search results after giving the search string mentioned above. The search was conducted using the title, keywords, or abstract fields depending upon the searching database. Although this area has papers from 1990, we have focused only from 2010 onward to provide more recent advances in the field. So, the search was conducted for the years ranging from 2010 to 2020.

B. INCLUSION AND EXCLUSION CRITERIA

Research studies to be included for this SLR must have related findings. They could be papers on application domains or the details of the development of tools or frameworks for information extraction. They must be peer-reviewed and written in English. The inclusion and exclusion criteria applied to filter the obtained studies are shown in Table 5.

C. STUDY SELECTION RESULTS

Figure 4. shows the Systematic Literature Review process followed to get the final study selection results by applying inclusion and exclusion criteria. From the primary keyword search string, 582 possible studies were identified from various databases, as mentioned in Table 4. Based on the keyword relevance, titles, and abstracts, these studies were screened, and relevant studies were grouped and duplicate, and irrelevant studies were removed. Subsequently, 105 relevant studies were selected from various databases, as outlined in Table 4. Afterwards, the references of all the selected studies were scanned to find the additional significant studies (that is snowballing). The aim here is to check that no study would be missed out during our search process. 15 additional

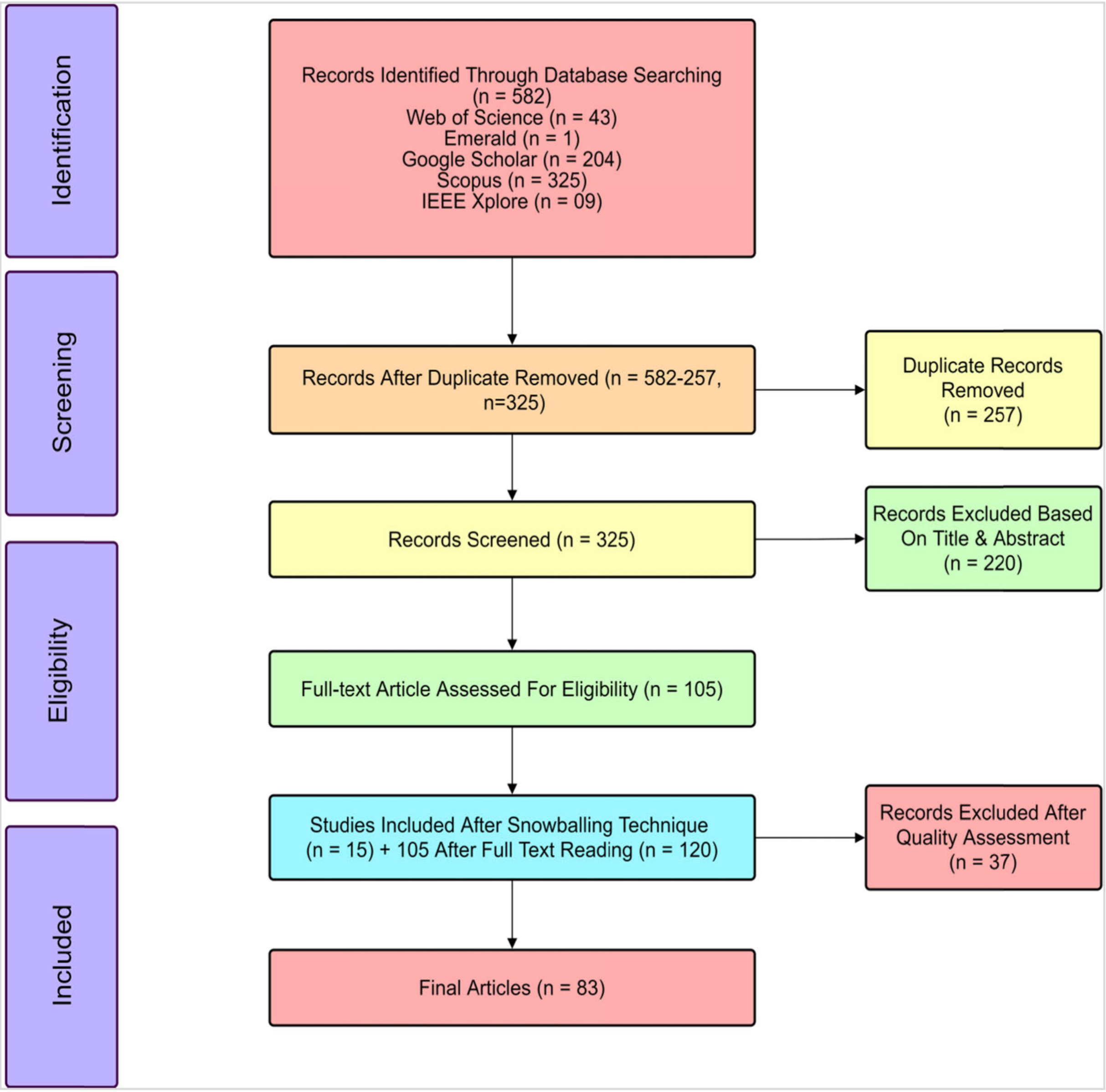

Identification
Screening
Eligibility
Included
Records Identified Through Database Searching (n = 582) Web of Science (n = 43) Emerald (n = 1) Google Scholar (n = 204) Scopus (n = 325) IEEE Xplore (n = 09)
Records After Duplicate Removed (n = 582-257, n=325)
Duplicate Records Removed (n = 257)
Records Screened (n = 325)
Records Excluded Based On Title & Abstract (n = 220)
Full-text Article Assessed For Eligibility (n = 105)
Studies Included After Snowballing Technique (n = 15) + 105 After Full Text Reading (n = 120)
Records Excluded After Quality Assessment (n = 37)
Final Articles (n = 83)

FIGURE 3. Flowchart for the selection of relevant papers.

TABLE 4. Literature databases search results.

Source Databases Searched	Count of studies obtained	Count of selected studies based on inclusion/exclusion criteria, removal of duplicate studies
Web of Science	43	05
Emerald	01	01
Google Scholar	204	15
Scopus	325	75
IEEE Xplore	09	09
Total	582	**105**

studies were obtained by snowballing. The selected studies count was then 120. Lastly, the quality assessment criteria to these 120 studies were applied. As a result, 83 research studies are finally selected.

The document type per year of selected studies is shown in Figure 5. Figure 6. shows percentage-wise contribution of types of studies.

D. QUALITY ASSESSMENT CRITERIA FOR THE RESEARCH STUDIES

Quality assessment permitted an evaluation of the significance of the studies to answer the research questions. Table 6. shows quality assessment criteria with a “quality score” of “4”. Research studies fulfilling this quality score were selected to conduct SLR.

TABLE 5. Inclusion and exclusion criteria for the research studies.

Criteria Number	Topic	Inclusion criteria	Exclusion criteria
1	Unstructured data/document analysis/document processing	The paper must focus on the challenges of the information extraction techniques to deal with the unstructured data.	Papers focussing on other unstructured data forms such as audio, video, and emails were excluded.
2	Unstructured dataset	The paper must contain details related to the name, size of dataset used, source of dataset, publicly available or self-built dataset used.	The papers lacking in information on data pre-processing techniques were excluded.
3	Techniques used for information extraction along with the identification and classification of named entities for the unstructured documents	The paper must contain information related to information extraction techniques and Named Entity Recognition for unstructured documents.	Blogs, online books, industry white papers on information extraction were excluded.
4.	Year range	Related papers that are published from 2010 to 2020	Grey papers which comprise of missing bibliographic information data like date and type of publication, issue numbers and volume, were excluded.
5.	Research question relevance	Papers that answer at least one research question.	The papers without any information related to research questions were excluded.

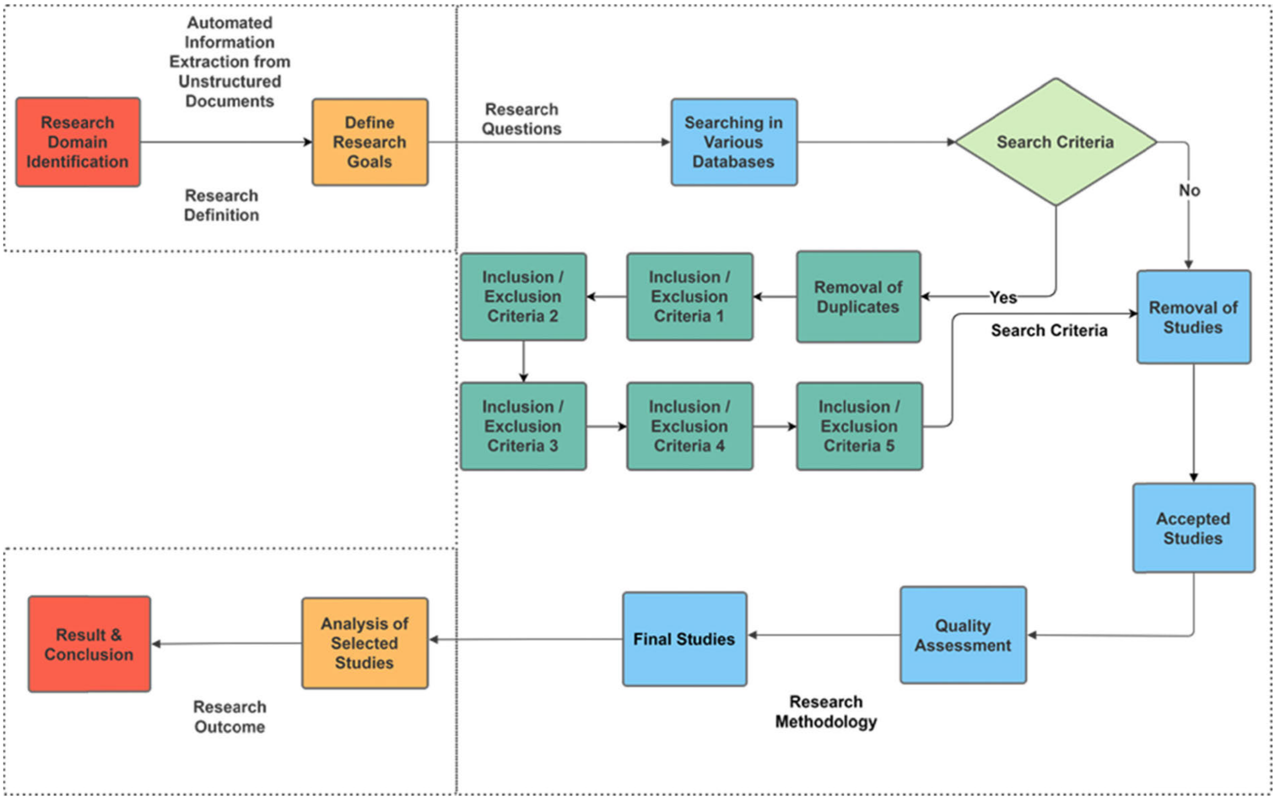

Automated Information Extraction from Unstructured Documents
Research Domain Identification
Define Research Goals
Research Definition
Research Questions
Searching in Various Databases
Search Criteria
No
Yes
Removal of Duplicates
Inclusion / Exclusion Criteria 1
Inclusion / Exclusion Criteria 2
Inclusion / Exclusion Criteria 3
Inclusion / Exclusion Criteria 4
Inclusion / Exclusion Criteria 5
Search Criteria
Removal of Studies
Accepted Studies
Quality Assessment
Final Studies
Research Methodology
Analysis of Selected Studies
Result & Conclusion
Research Outcome

FIGURE 4. Systematic Literature Review (SLR) process.

E. DATA EXTRACTION

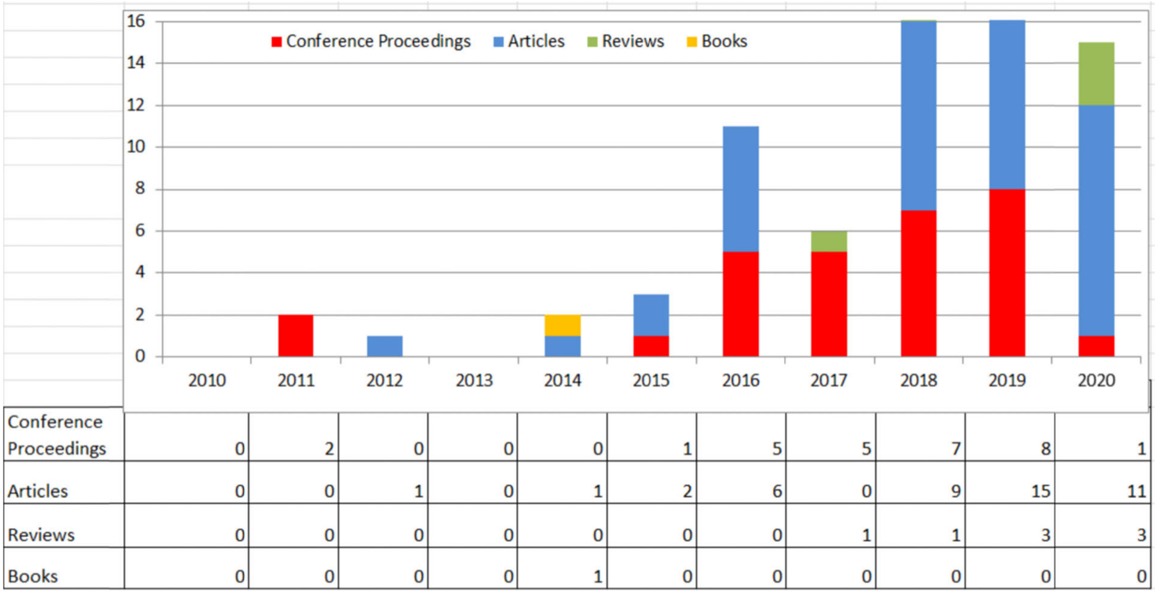
Table 7. shows a brief overview of the extracted data from selected studies based on their categorization to meet the goal of answering our research questions.

F. DATA SYNTHESIS

Figure 7. shows taxonomy of studied literature synthesized to answer research questions in detail. The green circle represents the advantages, and the red circle represents the disadvantages of the approach studied.

IV. BACKGROUND

This section will provide the necessary background information on all the AI-based approaches shown in Figure 7. Section 5 will then cover the literature on these three approaches.



	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
Conference Proceedings	0	2	0	0	0	1	5	5	7	8	1
Articles	0	0	1	0	1	2	6	0	9	15	11
Reviews	0	0	0	0	0	0	0	1	1	3	3
Books	0	0	0	0	1	0	0	0	0	0	0

FIGURE 5. Year-wise type of studies published.

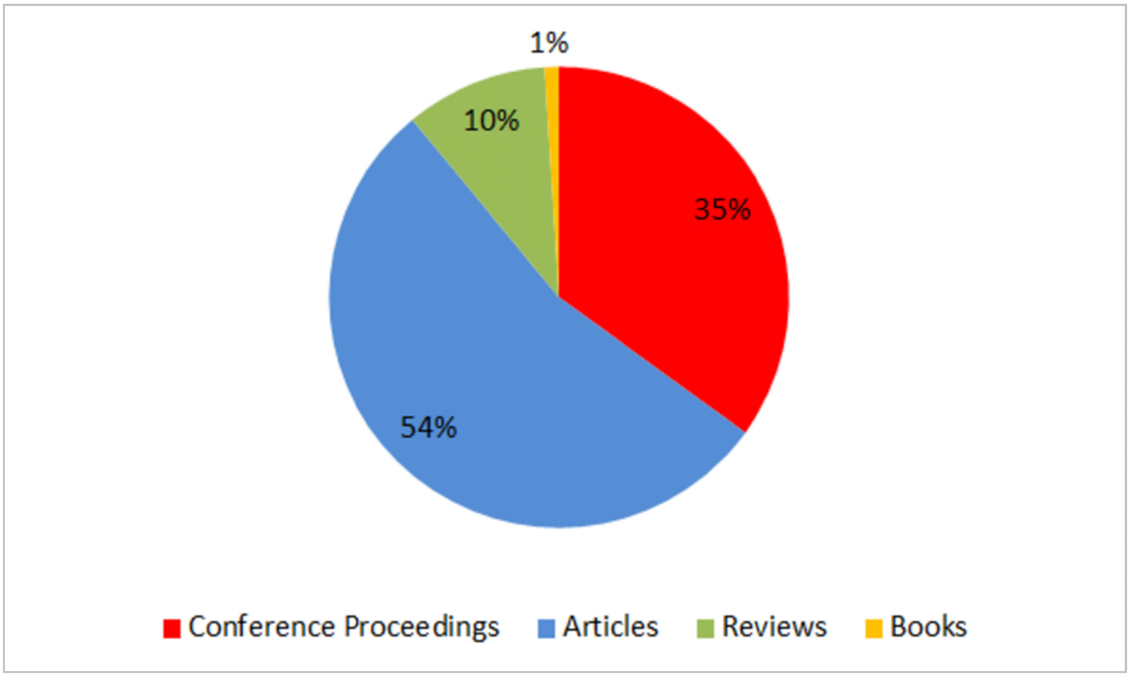


FIGURE 6. Percentage-wise contribution of types of studies.

A. OPTICAL CHARACTER RECOGNITION (OCR)

The manual extraction of text from the unstructured documents such as scanned PDF is not scalable and error-prone, as humans tend to get tired and make mistakes. The organizations have recently tried to use template-based approaches such as OCR to automate the document processing. OCR is used to recognize the text within an image; usually, a scanned printed or handwritten document. OCR can also automatically sort various document types and organize them according to the particular rules. For instance, classifying and managing invoices based on the type of product or vendor.

Figure 8. shows the main steps in OCR as discussed below:

Dataset [16] –Mostly, the researchers have used the publicly available datasets for handwritten or printed character recognition. The self-built dataset can also be used to extract the text using OCR.

Pre-processing [17], [18] –Pre-processing phase is needed to separate a character/word from the background in an image. It includes:

TABLE 6. Quality assessment criteria.

Criteria	Score

The studies must provide findings and results.	If criteria is satisfied
	Score=1
	Otherwise=0

The study must provide empirical proof on the findings.	If criteria is satisfied
	Score=1
	Otherwise=0

Are the research objectives and outcomes well presented in the paper?	If criteria is satisfied
	Score=1
	Otherwise=0

Are the references used in the research study are appropriate and sufficient?	If criteria is satisfied
	Score=1
	Otherwise=0

TABLE 7. Categorization of selected studies to answer the formulated research questions.

Category	Related data extracted from all selected papers
Techniques used for information extraction from unstructured data	The challenges with the existing information extraction techniques to deal with unstructured data.
Unstructured dataset and validation techniques	Publicly available unstructured document dataset. Data validation methods used to assess quality of data.
AI approaches used for unstructured data processing	Optical Character Recognition (OCR) - used for digitizing scanned document images, Robotics Process Automation (RPA) - used for automating a rule-based, repetitive task, other AI-based approaches such as RNN, CNN, Bi-LSTM, and BERT used for the information extraction from the unstructured documents.
Tools or frameworks developed by the researchers for information extraction from unstructured documents	Any details on tools or frameworks used/developed by the researchers mentioned in the selected studies.

- **Binarization:** It converts an image into a black and white pixel. This conversion is done by fixing a threshold value. If the value is greater, it is considered a white pixel else black pixel.
- **Noise Reduction:** It cleans the image by removing all the unwanted dots and patches.
- **Skew Correction:** Some text might be miss-aligned. Skew correction helps to align this text.
- **Slant Removal:** Some images in the dataset can have slant text; this technique is used to remove the slant from the text.

Segmentation [19] – It breaks an image into parts for further processing, which can be segmented as text or word form. Text line detection and word extraction methods are used for segmentation. Correct recognition of character depends on the accurate segmentation.

Feature extraction [20] – It extracts the raw data into manageable data or required data. Two widely used feature extraction methods in OCR are: statistical (identifies a statistical feature of character) and structural (identifies structural features like horizontal and vertical lines, endpoints).

Statistical feature extraction can be done by zoning, where images are divided into the zones, and then the features are extracted from each zone to form the feature vector.

Projection & profile can also be used as a statistical feature extraction method. Projection histograms calculate the number of pixels in a different direction of a character image. It can be used to separate the characters such as ''m'' and ''n.''

A profile is used to calculate the number of pixels from the bounding box to the outer edge. It is used to define the external shapes of the characters. It allows distinguishing between the letters, such as ''p'' and ''q.''

In structural feature extraction, the geometrical properties of a symbol or character are extracted. The geometrical properties or the structural features of the character are - character strokes, horizontal lines, vertical lines, endpoints, intersections between lines, and loops. It provides the idea about the type of components, that makes up the character.

Classification – In most OCR techniques, an algorithm learns to categorize or classify the character set and numerals accurately, as it is trained on a known dataset. The most popular techniques used for the classification in OCR literature studies are mentioned below:

- **K Nearest Neighbor:** It classifies the objects with a similar feature in close proximity. It is used to segment, and recognize Latin alphabets in uppercase and lowercase, and Devanagari consonants and vowels for Indian scanned document images of Latin and Devanagari scripts in the study [20].

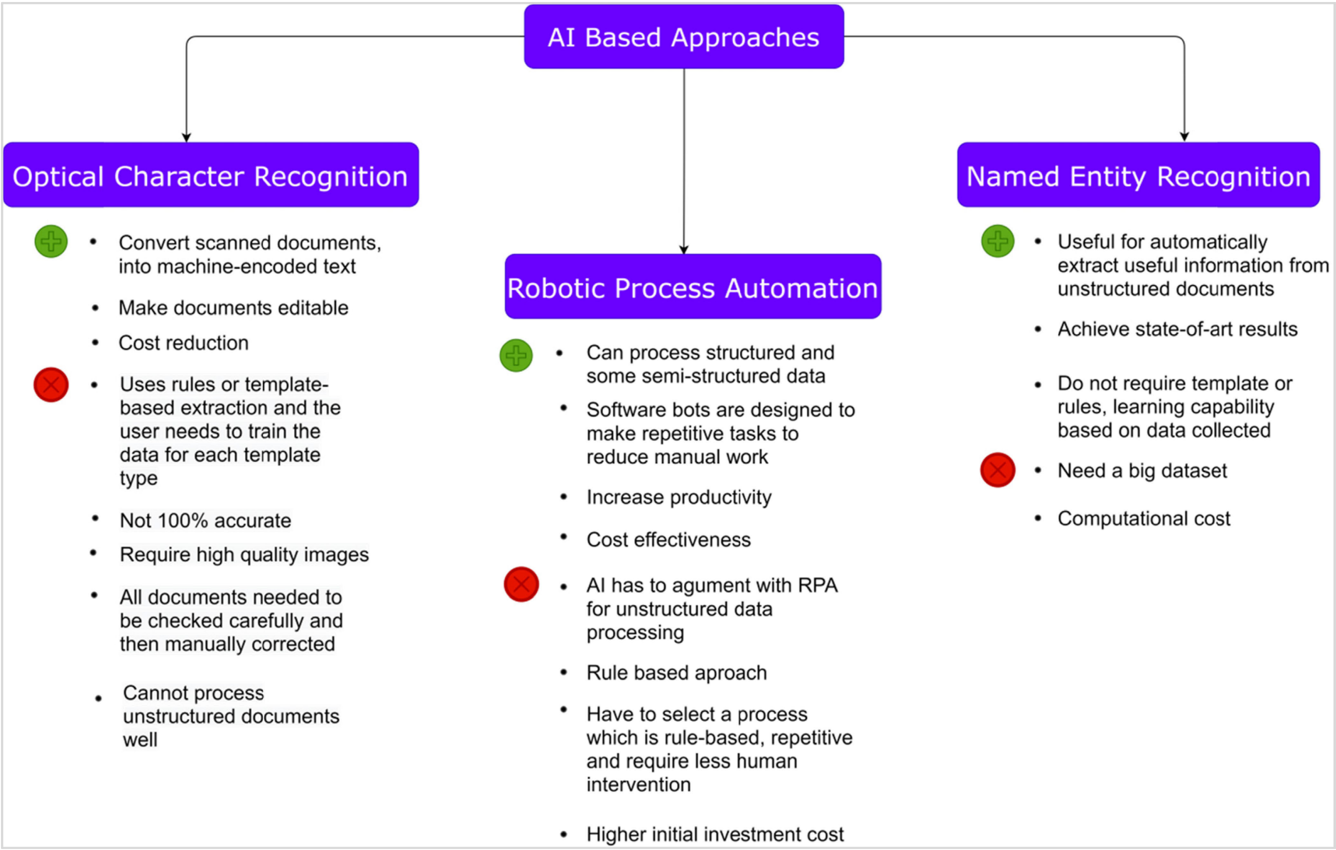

AI Based Approaches
Optical Character Recognition
Convert scanned documents, into machine-encoded text
Make documents editable
Cost reduction
Uses rules or template-based extraction and the user needs to train the data for each template type
Not 100% accurate
Require high quality images
All documents needed to be checked carefully and then manually corrected
Cannot process unstructured documents well
Robotic Process Automation
Can process structured and some semi-structured data
Software bots are designed to make repetitive tasks to reduce manual work
Increase productivity
Cost effectiveness
AI has to agument with RPA for unstructured data processing
Rule based aproach
Have to select a process which is rule-based, repetitive and require less human intervention
Higher initial investment cost
Named Entity Recognition
Useful for automatically extract useful information from unstructured documents
Achieve state-of-art results
Do not require template or rules, learning capability based on data collected
Need a big dataset
Computational cost

FIGURE 7. Literature taxonomy.

- **Naïve Bayes Classifier:** It is a probabilistic classification method. It uses the Bayes theorem of probability to calculate the class of new or unknown data. It is used in business invoice recognition and classification of fields like invoice_number and invoice_date in [21].
- **Neural Network:** It has shown strong abilities to automate text detection and data extraction by recognizing the underlying relationships of characters/words. Region Based Convolutional Neural Networks (R-CNN) is used for the object detection in real-world Chinese passport and medical receipt dataset in [22].
- **Support Vector Machine:** It is the commonly used classification algorithm in OCR. It performs better than any other classification method. It is not based on any assumptions of independence as in the Naïve Bayes method. It is used for the text categorization or recognition in [20], [23].

Post-processing [24] – It detects and corrects the misspelling in the output text after an image is processed using the OCR technique.

Metrics [24] – Metrics are used to calculate word and character error rates and to evaluate the performance of OCR techniques.

B. ROBOTICS PROCESS AUTOMATION (RPA)

Another method to automate the unstructured document processing is to use a rule-based method called as Robotic Process Automation (RPA). This method requires humans to write simple rules to perform repetitive actions with a software bot.

Figure 9. shows the thematic diagram of RPA as discussed below:

Automated software robots [25]: RPA allows the organizations to automate the highly redundant and rule-based tasks at a small amount of the earlier incurred price and time. It is scalable. It can perform many tasks such as logging into the system or application, placing files and folders at the selected locations, data copying and pasting, form filling, extracting structured data from the documents, web browser scraping, and more. These software robots increase the scalability, speed and reduce the operational cost.

Cognitive RPA [10]: Cognitive RPA takes advantage of NLP and ML to do high-volume repetitive tasks previously performed by humans. Traditional RPA may not be enough to automate the systems to make the simple decisions independently. This is where cognitive RPA plays an important role. It is a combination of RPA and AI. It can manage huge data, find hidden patterns from the data, and predict the future trends. It improves the performance over time. It can imitate the way a human thinks and can make intelligent decisions. It can process unstructured documents and can be integrated with the analytics tool of the organization or Business Process Management (BPM) applications.

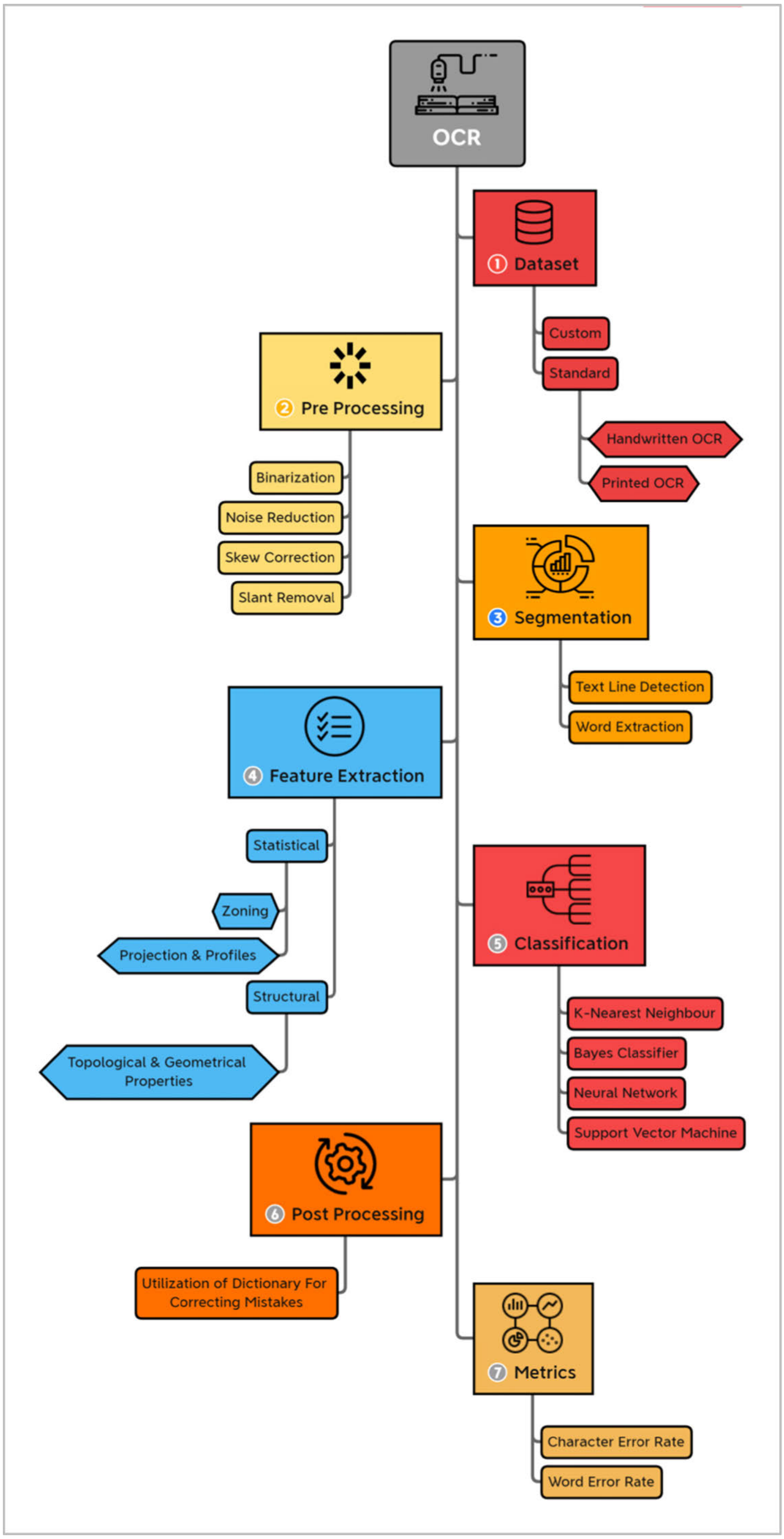


FIGURE 8. Main Steps in OCR.

Process selection as RPA candidate [26]: One of the important facts that possibly impact RPA implementation success is the suitability of the candidate process for automation. The organizations need to know the process appropriateness criteria for implementing RPA. The selected process can be first defined into clear or definite rules, as RPA is only

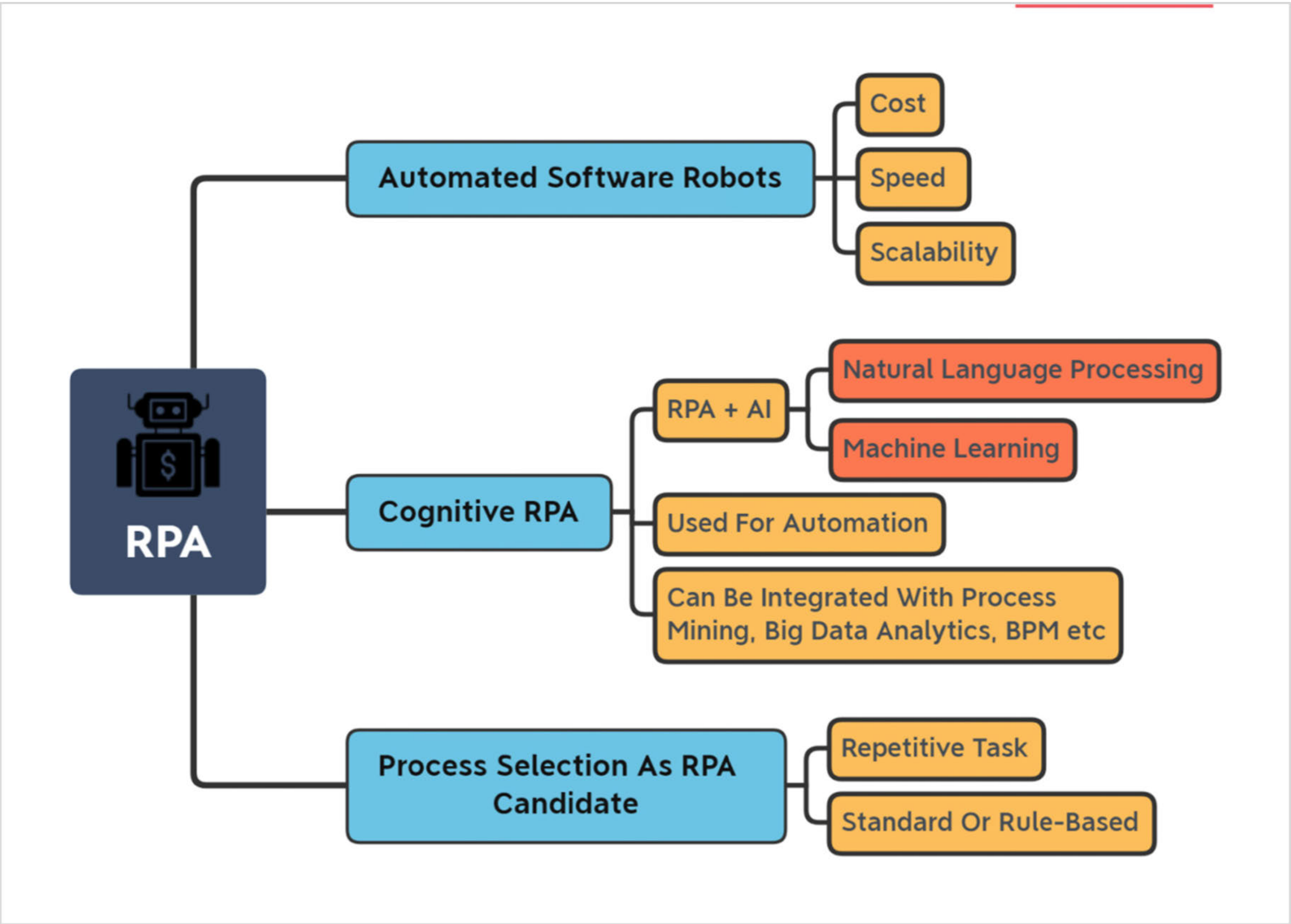

RPA
$
Automated Software Robots
Cost
Speed
Scalability
Cognitive RPA
RPA + AI
Natural Language Processing
Machine Learning
Used For Automation
Can Be Integrated With Process Mining, Big Data Analytics, BPM etc
Process Selection As RPA Candidate
Repetitive Task
Standard Or Rule-Based

FIGURE 9. Thematic diagram for robotics process automation for unstructured document processing.

suitable for the rule-based tasks. The process standardization before automation is also essential, as the more standardized process causes less exceptions.

Figure 10. shows the main steps in RPA as discussed below:

1. **Identify the processes to automate** [26]: The organizations first need to identify the process that is appropriate for automation.
2. **Automate the processes using software** [10]: The tasks performed by the identified process is defined in terms of series of instructions. Process automation workflow is defined in this step.
3. **Developing a bot controller** [25], [27]: Automated processes are pushed to a bot controller. The bot controller is the most important step of any RPA workflow. It is used to control and prioritize the process execution. It allows users to schedule, manage, and control various activities. The bot controller also controls the process execution status by checking execution logs.
4. **Integration with enterprise application** [25]: Automated software bots can be integrated with the analytics tool of the organization or Business Process Management (BPM) applications.

C. NAMED ENTITY RECOGNITION AND OTHER ARTIFICIAL INTELLIGENCE-BASED APPROACHES

AI-based approaches have a strong potential to extract useful information from unstructured documents automatically. Figure 11. shows the thematic diagram for AI-based approaches as discussed below:

Dataset [28]: Various researchers use the publicly available datasets popularly for the document analysis tasks. To get the insights from the scanned documents such as receipts, invoices, researchers have prepared their datasets or worked on the dataset provided by the specific organizations. The details on the datasets are discussed in Section V.

Feature extraction [29]: Various feature extraction methods used in the text classification and recognition are: GloVe, TF-IDF, and Word2Vec. The GlobalVectors (GloVe) is used to get the vector or numerical representation for words. TF-IDF is a popular approach to assign the weights to words, that indicates their importance in the documents. Word2Vec is the most standard method to learn word embeddings from considerably large datasets using Neural Networks (NN). Word2Vec can be performed using Continuous Bag of Words (CBOW) or Continuous Skip Gram.

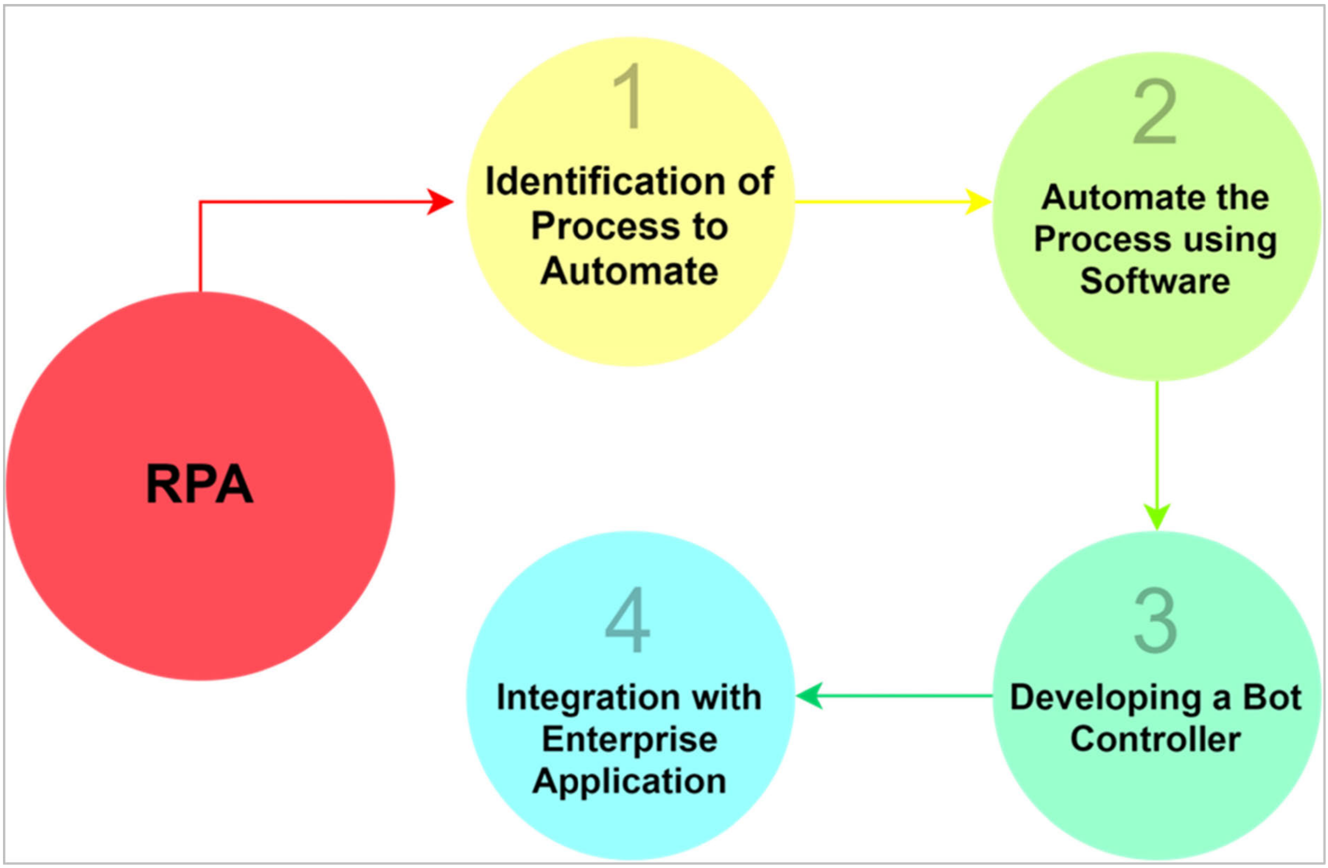

RPA
1
Identification of Process to Automate
2
Automate the Process using Software
3
Developing a Bot Controller
4
Integration with Enterprise Application

FIGURE 10. Main steps in RPA.

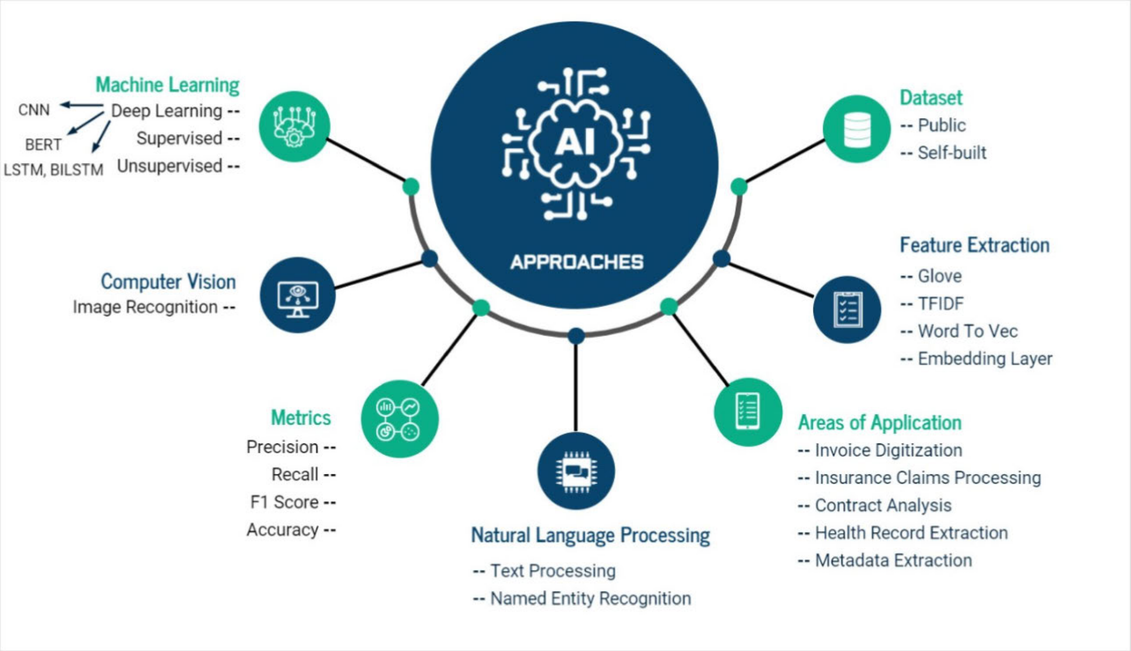


FIGURE 11. Thematic diagram for AI-based approaches.

Areas of applications [6]: AI-based approaches are used in various applications such as invoice digitization, health record extraction, metadata extraction, insurance claims processing, contract analysis, and many more.

Natural Language Processing [23]: It analyses the grammatical structure at the sentence level and then creates grammatical rules to obtain the useful information about the sentence structure. Among all the techniques, NER tech-

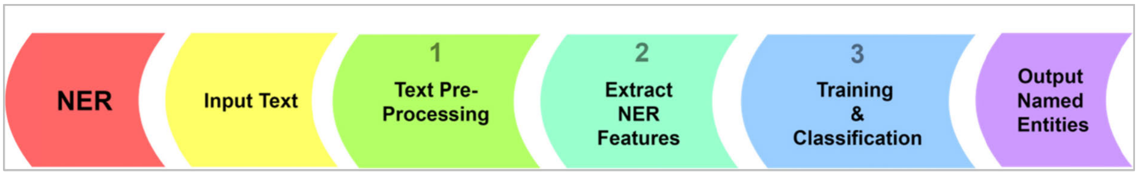


FIGURE 12. Main steps in NER.

niques serve the most basic and essential techniques in NLP. NLP uses the sentence-level syntactic rules such as assigning grammar rules and patterns at the word or token levels, such as the regular expressions for information extraction from the given text, using NER. It automatically scans the unstructured text to locate the "named entities" like a name (first name, last name), location (such as countries, cities), organization, date, and invoice numbers in the text [30].

Figure 12. shows the main steps in NER as discussed below:

1. **Text pre-processing**: It transforms the given text into a format that ML algorithms can understands better. It includes: Tokenization, normalization, and noise removal. Tokenization is splitting the text into smaller components, referred as "tokens". Normalization removes the stop words, and convert all text to lowercase characters. Noise removal performs text cleaning by removing extra white spaces.
2. **Extract NER features:** NLP model cannot work on the raw text data directly. So, feature extraction methods are required to convert text into a numerical representation of features or a matrix (or vector) of features.
3. **Training and classification:** The extracted features are passed through a NER model, that will classify different words and phrases into specific categories.

Metrics [31]: Different metrics, for example, precision, recall, and F1 score are used for the model evaluation. Precision discusses the prediction accuracy of the model by calculating the number of actual positives out of the total predicted positive. If the rate of False Positives is high, precision is a good measure to use.

$$Precision = \frac{True\ Positive}{Total\ Predicted\ Positive}$$

Recall calculates the number of the actual positives, a model can capture. If the rate of False Negative is high, recall is a good measure to use.

$$Recall = \frac{True\ Positive}{Total\ Actual\ Positive}$$

F1 score is a suitable metric to seek a balance between the precision and recall. When the rate of Actual Negatives is high, that is, for uneven class distribution, the F1 score is a good measure to use.

$$F1score = 2\frac{Precision \times Recall}{Precision + Recall}$$

Computer Vision [32]: Few researchers also focused on the Computer Vision approach for identifying the scanned document images, by defining a bounding box over the interesting image area, which is to be extracted.

Machine Learning [32]–[34]: The studied approaches were classified into three categories of the techniques, which are: Deep Learning, Supervised Learning, and Unsupervised Learning. Automatic feature learning from the given data is an important characteristic of Deep Learning (DL) models. It is their biggest advantage, and it is referred to as the feature learning. Defining a suitable Neural Network (NN) model and providing the right labeled dataset is sufficient in Deep Learning. During model training, the network tries to learn and extract the useful, accurate features from the data. Convolutional Neural Networks (CNN) is an example of a Deep Learning network mostly used in the text classification and recognition field. It has the excessive capability to capture the local features that provide an excellent help to the researchers, who analyze and utilize image data. Bidirectional Encoder Representations from Transformers (BERT) is a language model used in NLP. BERT can capture most of the local and global feature representations of a sequence of text. Bi-directional Long Short Term Memory (Bi-LSTM) is combining two independent LSTM. This arrangement permits the networks to keep both reverse and forward data, about the text sequence every time.

In Supervised Machine Learning approaches, the model learns from the historical or past data and uses that knowledge learning to the existing data to forecast the future events. Text categorization is a Supervised Machine Learning example.

In Unsupervised Learning, the model trains on the data that is neither categorized nor labeled. It means without providing the labeled data, the Machine Learning model learns by itself. The Unsupervised Learning model needs to categorize the data without any prior knowledge about the data. Fraud detection is an example of the Unsupervised Learning approach.

Table 8. provides the summary of approaches and their example usage in automatic information extraction from unstructured documents.

V. RESULTS

This section will outline the results for each of our research questions. We first outline the key issues and challenges faced during the unstructured document processing.

A. RQ1-CHALLENGES WITH INFORMATION EXTRACTION TECHNIQUES TO DEAL WITH THE UNSTRUCTURED DATA

Enterprises have a sophisticated system to record and utilize the structured data either using the excel or an enter-

TABLE 8. Summary of approaches and their example usage in automatic information extraction from unstructured documents.

Approach used	Example use in the unstructured document processing	References
OCR	• Handwritten and printed scanned image document text retrieval, recognition, and classification • Invoice classification and extraction • Receipt digitization	[22], [24], [35]–[39],[21]
RPA	• Automation for healthcare staffing services • Automation for banking & financial services • Automation for property & casualty insurance • Automate HR activities in the company	[40],[41]
NER	• Legal clause extraction • Invoice extraction • Receipt extraction • Clinical information extraction • Scientific article metadata extraction	[42]–[46],[21], [36], [47], [48],[13], [29], [49], [50],[36], [51], [52]

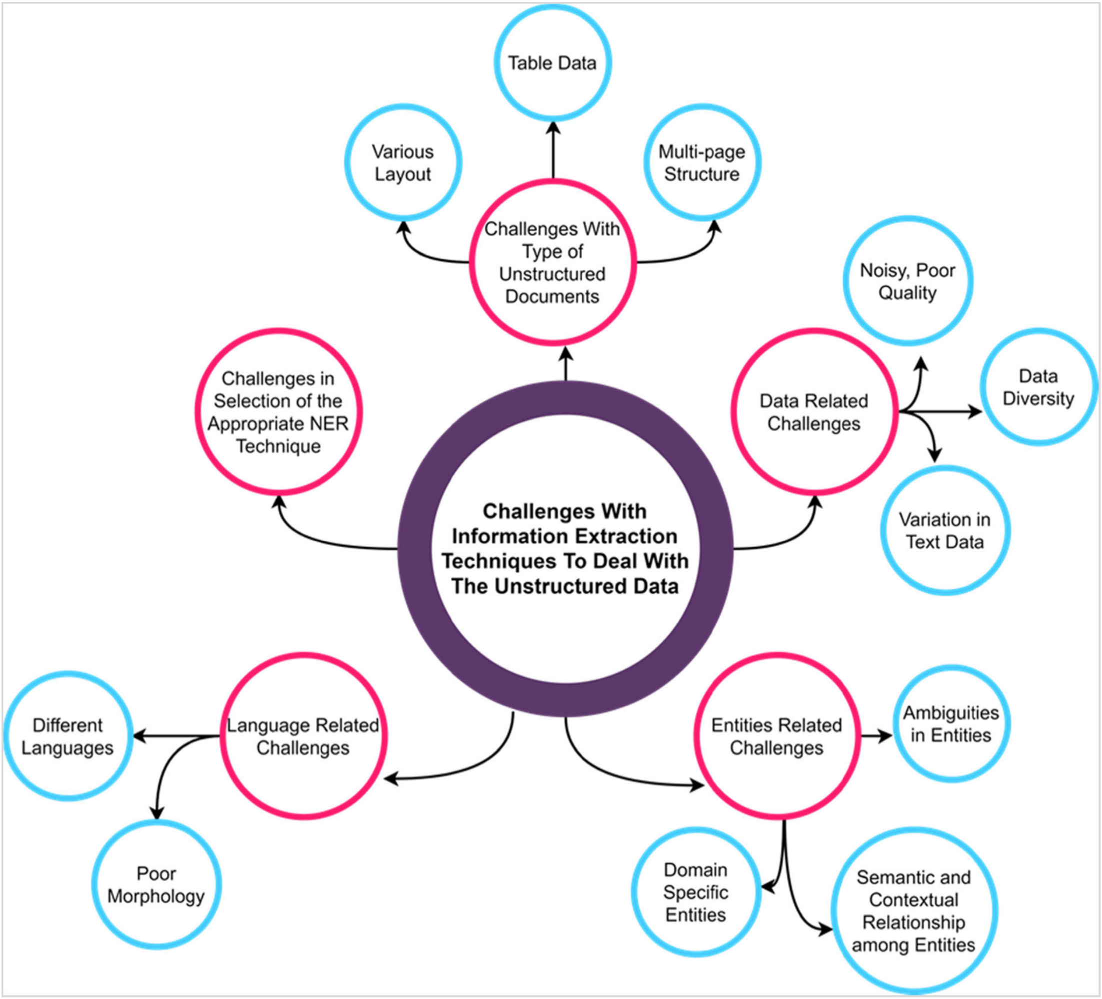


Table Data
Various Layout
Multi-page Structure
Challenges With Type of Unstructured Documents
Noisy, Poor Quality
Data Diversity
Challenges in Selection of the Appropriate NER Technique
Data Related Challenges
Challenges With Information Extraction Techniques To Deal With The Unstructured Data
Variation in Text Data
Different Languages
Language Related Challenges
Entities Related Challenges
Ambiguities in Entities
Poor Morphology
Domain Specific Entities
Semantic and Contextual Relationship among Entities

FIGURE 13. Challenges with information extraction techniques to deal with the unstructured data.

prise database system. However, a much larger proportion of enterprise data these days is the unstructured data [2]. The unstructured data has no pre-defined schema or format, making it much more challenging to collect, process, and analyze. It is more challenging to analyze since it lacks the proper structure [1].

We identified several factors that affects the performance of the information extraction techniques. We studied the challenges based on unstructured data, named entities involved, domain, and language-related limitations [6].

Figure 13. shows the challenges faced by the information extraction techniques when dealing with the unstructured data.

1. **Data related challenges:** The unstructured data faces major data-related challenges. Sometimes OCR text extraction adds incorrect text data in the form of noisy data. This is a major data related challenge [24]. The unstructured data generated from multiple sources is non-standard and has different data formats. It is called data diversity. Variation in the text in unstructured text document can also be a major issue for the traditional information extraction techniques [53]. Lack of enough data and poor-quality data is another challenge in unstructured data processing [6].
2. **Entities related challenges:** Extracting information from a highly ambiguous language such as Arabic, especially without creating a dictionary, is challenging. The semantics and the contextual relationship among named entities for such ambiguous language are challenging for the information extraction techniques [6]. Domain specific entities poses another challenge. For example, domain specific entities of the biomedical datasets differ from any other domain dataset [6], [49], [54].
3. **Language related challenges:** Poor morphological languages add a challenge to the information extraction [55]. Information extraction techniques varies for different languages [6].
4. **Challenges in selection of the appropriate NER technique:** Selection of the appropriate NER technique for extracting the named entities depends on the language and the domain. The lack of a large labeled corpus is another challenge. Creating such a huge and manually labeled data is a time-consuming and tedious task [1].
5. **Challenges with type of unstructured documents:** The business process needs to process different unstructured documents given by the client or the supplier, such as invoices, passport data, ID-card, various application forms, and much more. All these unstructured documents are of a different type, form, and layout by nature. The information extraction techniques should classify these documents by their type and nature. It should extract the required fields from the particular document by applying either a template-based or a template-free approach [33]. Another challenge for the information extraction techniques is to process and enhance the quality of the scanned documents, as the documents submitted by the client or supplier are generally scanned with a low-quality scanner or mobile devices. Multi-page unstructured documents consisting of tables with data spanning across different pages complicate retrieval of the correct target data from the document [56].

B. RQ2-DATASET

This section will cover some of the most widely used publicly available datasets for information extraction from the unstructured documents. We also surveyed the key issues and challenges with existing datasets (RQ2).

Data is the foundation of any AI-based model. Obtaining correctly sourced and relevant contextual data and checking for data bias, will help to build a better performance model.

1) PUBLIC DATASETS FOR THE UNSTRUCTURED DATA

We observe that researchers have used different datasets to train the model for specific information extraction tasks, or the unstructured document analysis tasks. Getting the right dataset containing sufficient quantity and quality data for AI-based model training and testing, is essential to achieve good research results.

The earlier research in OCR in several diverse languages, like, English and Arabic, has focused on the publicly available datasets such as the MNIST dataset [57]. The modified NIST(MNIST) is the most used/cited datasets for handwritten digit recognition in English language. It is the subgroup or part of the NIST dataset, and so it is termed as a modified NIST or MNIST. MNIST samples are normalized. Thus, it reduces the data pre-processing and structuring time. It has 60,000 training sample images and 10,000 testing sample images.

OCR for the Arabic language uses PAWs (Printed Arabic Words) dataset [58]. PAWs have all the words in Arabic language. Arabic words consist of one or more sub-words (PAWs). It contains 83,056 PAWs images which are extracted from 5,50,000 diverse words. Every word sample image is collected in five different font styles— Naskh, Thuluth, Kufi, Andalusi, and Typing Machine. It is used for document image analysis and recognition tasks for Arabic input.

The systematic literature review [16] on handwritten OCR has provided the summary of publicly available datasets for Chinese, Indian, Urdu, and Persian/Farsi languages, for example, CEDAR (English), CENPARMI (Farsi), PE92 (Korean), UCOM (Urdu), HCL2000 (Chinese).

In 2002, the University of Buffalo has built the dataset called Center of Excellence for Document Analysis and Recognition (CEDAR). It is an online handwritten text dataset for the English language, consisting of text lines, written by approximately 200 writers and stored in an online format.

The CENter for Pattern Recognition and Machine Intelligence (CENPARMI) presented the initial version of the Farsi dataset in year 2006. It includes 18,000 examples of Farsi numerals, consisting of 11,000, 2,000, and 5,000 samples for training, verification and testing purposes respectively.

TABLE 9. Summary of publicly available datasets, their application domains, and their feature.

Reference paper	Dataset Name	Application & Purpose	Feature
[61]	Ryerson Vision Lab Complex Document Information Processing (RVL-CDIP)	Dataset for OCR and additional metadata, used for the document classification and analysis.	It consists of 4,00,000 grayscale images for 16 different classes, with 25,000 images per class.
[62],[28], [63]	Medical Information Mart for Intensive Care (MIMIC) database and Informatics for Integrating Biology and the Bedside (i2b2)	Healthcare dataset used to extract symptom and disease.	It is an open-access intensive care database in clinical narrative form.
[64],[51]	GROund Truth for Open Access Publications (GROTOAP)	Scientific article dataset for metadata extraction, used for zone classifiers feature selection and SVM parameters determination.	It consists of 113 documents (TrueViz format) consisting of the text content along with their geometric features and zone labels.
[48]	International Conference on Document Analysis and Recognition Scanned Receipts OCR and Information Extraction (ICDAR-2019 SROIE) dataset	Receipt dataset, used for receipt details extraction.	The dataset includes 1000 printed, scanned receipt images. Each receipt image comprises of main text fields: item name, item price, and total cost of items. Digits and English characters are annotated in the dataset.
[65]	Document deskewer dataset	Optical Character Recognition, used for image document skew detection.	It includes the documents used for deskewing in OCR.
[66], [67], [68]	Internet Movie Database (IMDB)	Movie Reviews dataset, used for the classification of movies in negative & positive class label.	It includes 100K movie reviews with no more than 30 reviews per movie.
[66],[69]	Yelp	Review dataset, used for classification of electronic product and few other category reviews.	It includes information on business reviews, users, businesses, photos.
[66],[70]	20news	Newsgroup document dataset used for text document classification & clustering	It includes multiple classes of 20K newsgroup documents with 20 nominal document categories, five each related to a distinct topic.

PE92 (collected by POSTECH funded by ETRI) is a handwritten Korean character image dataset, consisting of 100 sets of KS 2350 Korean handwritten character images. It has a collection of different writing styles.

The Comprehensive handwritten dataset for Urdu language (UCOM), an Urdu language dataset, is used to recognize the characters and writer identification. It contains 53,248 character images and 62,000 word images written in nasta'liq, that is, calligraphy style.

The Handwritten Chinese Language character-2000 (HCL2000) is a Chinese handwritten character dataset. It includes 3,755 most commonly used Chinese handwritten character images recorded by 1,000 distinct subjects. It is exclusive because it consists of mainly two sub-categories; one is a Chinese handwritten characters dataset, and the other category is the metadata of the corresponding writer information dataset. These categories of the dataset can be used for the character image recognition task and writer's metadata extraction tasks. For example age, gender, occupation, and education of a writer can be extracted.

The research study [57] has used the dataset Mobile Identity Document Video dataset (MIDV-500). It includes 500 video clips for 50 diverse ID types, containing 14 passports, 17 ID cards, 13 driving licenses, and six other identity documents of different countries with ground truth, which allows performing research in various document analysis problems.

Pattern Recognition and Image Analysis (PRImA) [59] dataset consists of the document images of several types and layouts used for printed document layout analysis.

Another dataset built upon the scanned color document images of multiple layouts is the UvA dataset [60]. UvA is a color document image analysis dataset used for the layout detection and segmentation task.

The dataset PubMed [13], [51] includes 26 million citations for biomedical literature papers from MEDLINE, life science journal articles, and online books. These references also contains links to the full-text content articles from PubMed Central and publisher websites. It also includes the bibliographic references of the documents with their metadata information. This feature makes the PubMed dataset the most valuable and popular dataset for the metadata extraction tasks.

Table 9. shows other widely used task-specific public datasets for automatic information extraction from unstructured documents.

2) CHALLENGES/ISSUES WITH EXISTING DATASETS

We explored and surveyed the literature from the context of the availability of the datasets, and we conclude that the existing dataset has several open challenges/issues. Figure 14. shows these challenges/issues with the existing datasets.

1. **Poor quality of datasets:** The major challenge observed in the existing dataset is its quality. Quality data is required for the information extraction models to function well. The scanned document images in most available datasets have low-resolution quality, leading to poor OCR results. Few images in datasets are off-centered or tilted, which are skewed images. The images having undesired or distorted information like the patterns or watermarks in the background can be classified as noisy images [39]. These datasets also have missing or omitted data values and few other errors, that gives less informative and meaningless extractions [35].

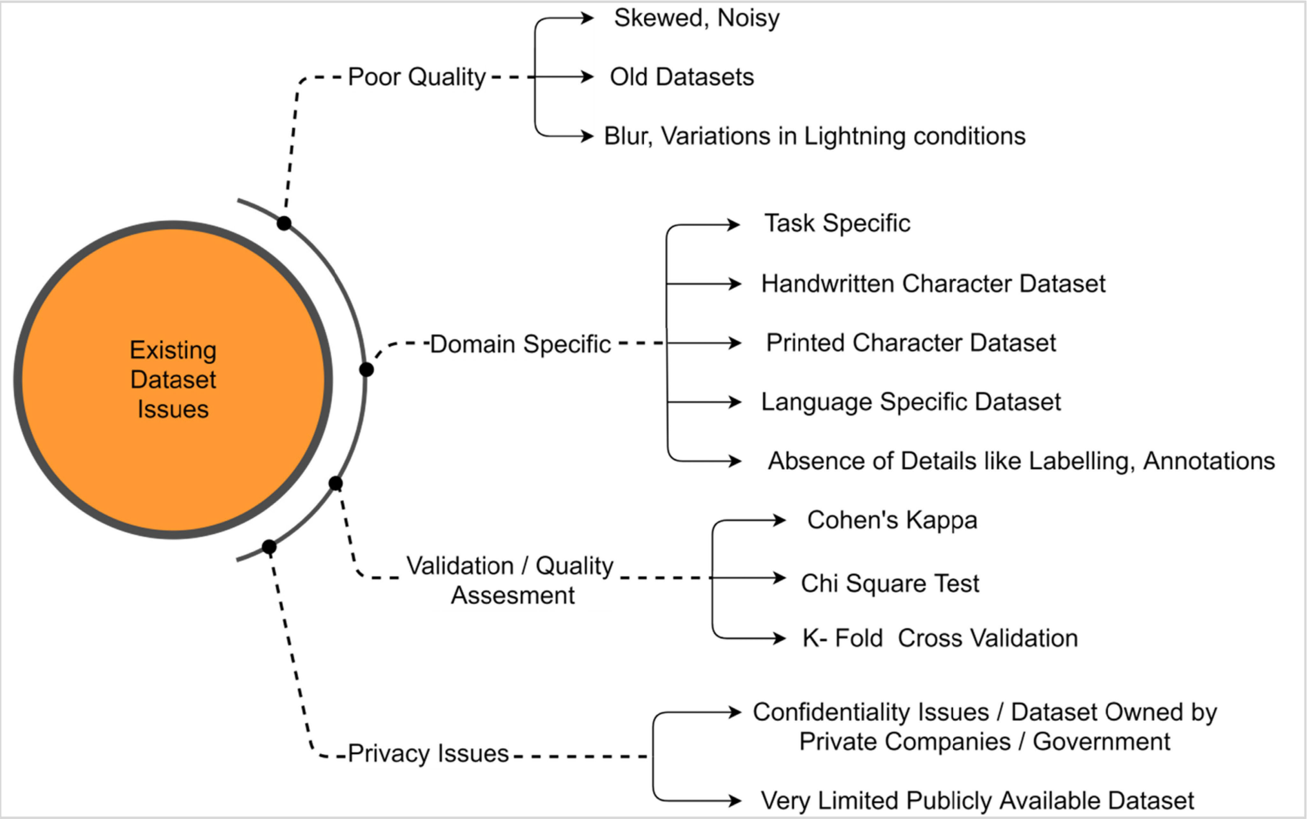

Existing Dataset Issues
Poor Quality
Skewed, Noisy
Old Datasets
Blur, Variations in Lightning conditions
Domain Specific
Task Specific
Handwritten Character Dataset
Printed Character Dataset
Language Specific Dataset
Absence of Details like Labelling, Annotations
Validation / Quality Assesment
Cohen's Kappa
Chi Square Test
K- Fold Cross Validation
Privacy Issues
Confidentiality Issues / Dataset Owned by Private Companies / Government
Very Limited Publicly Available Dataset

FIGURE 14. Challenges/issues in existing datasets.

Another challenge is that some datasets have outdated, non-standard contents. Old datasets like RVL-CDIP do not contain adequately scanned images, which can be an issue for the text extraction process. It also consists of some common problems like blurriness and variation in lightning conditions, which further complicates the image pre-processing [55].

2. **Domain specific datasets:** Existing publicly available datasets are very task-specific; that is, they are related to the data extraction of the scientific articles or clinical information that is not generalized [28]. In handwritten datasets, various kinds of handwritings are present, even cursive text, making it challenging for the OCR to detect and extract the actual text, leading to less accurate results [56]. In such cases, the advanced OCR techniques are needed. Few printed documents like the receipts and the invoices may contain handwritten remarks or characters or numerals, requiring advanced OCR techniques for the recognition [61]. Datasets are also available in different languages like Chinese passport and medical receipt dataset [22]. Language-ambiguity, poor morphology, language-dependent annotations, and the unavailability of large labeled and annotated corpus in the public dataset could be an issue [6], [55].
3. **Data Validation/quality assessment techniques:** We observed that very few research studies had reported the data validation techniques to check data quality [21], [28], [36], [62], [71], [72]. Quality assessment tests should be conducted on the datasets to check whether the available data is suitable or fit to train the model. Despite the usefulness of the data validation methods, one major issue is that the existing literature lacks a focused discussion on different data validation techniques that are available for categorical/nominal data and selection criteria of these techniques. Another challenge to explore the data validation techniques is that not many quality assessment techniques are discussed in the existing literature. The existing literature also lacks the discussion on the reasons behind choosing a particular data validation technique for their dataset. Few studies included Chi-square [73], [74], Cohen's kappa [28], [62], and K-fold cross-validation [21], [36], [71], [72] as a statistical measure of data quality assessment.
4. **Privacy issues:** Most unstructured datasets should have an extremely high computational power system to improve execution ability and decrease processing delay. Most self-built document datasets are confidential or sensitive. They contain private information about individuals, administrations, or companies. For example, an invoice is a private document to any organization because of which datasets related to invoices having public access are scarce. So, this kind of dataset is not publicly available [21], [72]. Several conferences and workshops provide custom-built datasets to their

participants for the document analysis and recognition purposes. The organizers usually keep the collected training and test data on the internet for the participants. Data is provided to the participants prior to such competitions. The delivered dataset is publicly accessible for research purposes. However, registration for such datasets is typically required. For example, ICDAR has such types of datasets, on which many researchers have proposed their work [33], [75].

C. RQ3-DATA VALIDATION TECHNIQUES

In this particular section, we surveyed the data validation techniques researchers have used for assessing the quality of data used to train their AI-based model for information extraction from unstructured documents.

Data validation refers to a method used to assess the quality of data used for training an AI-based model. Data validation is a laborious and time-consuming task, particularly if there is a large dataset and is aimed to perform the validation manually. Data validation is important as it guarantees that the analysis results are accurate, since the training model has precise or right data to solve the specific problem. Data validation provides high-quality data.

Reviewed studies reported significantly fewer data validation methods of the unstructured dataset, to assess data quality used to train AI-based models. These data validation methods include Chi-square, Cohen's Kappa, and K-fold cross-validation.

Chi-square is a commonly used feature selection method for data quality improvement, by measuring the dependency between a feature and a class label [73], [74]. A class label refers to a discrete feature, the value of which is wished to be predicted based on the values of other features. The Chi-square test helps to choose the most relevant features to train the model. It checks whether the input variables/features are dependent or independent of output variables. If variables are independent, then they are removed from dataset. It is a non-parametric test. It can only compare nominal/categorical, that is, non-numerical variables such as gender, name, or address. It is one of the most suitable statistical techniques for hypotheses testing, when the variables are nominal/categorical. Many information extraction tasks require nominal/categorical types of data, for example, name, address.

Another data validation method is Cohen's Kappa, generally used after manual annotations of data. Few research studies used self-built datasets, which were labeled manually [18], [35]. There are two possible limitations in such manual annotations: the labeling bias of the process (that is, annotations) and incorrect labeling. Annotating the data by more than one individual and performing validation methods to calculate the "inter-rater agreement" amongst the annotators could answer these two limitations. Few studies have adopted this validation method for the work of the annotator, which is Cohen's kappa [28], [62]. It is a statistical measure of inter-rater reliability/agreement denoted by "κ coefficient". It is the degree of agreement among the annotators. It is a score of how much similarity or agreement exists in the annotations given by various annotators. The example phenotype chosen by the annotator and their κ coefficient as inter-rater agreement measure is shown in [62]. The κ coefficient's strength can be interpreted as: 0.01-0.20 slight; 0.21-0.40 fair; 0.41-0.60 moderate; 0.61-0.80 substantial; 0.81-1.00 almost perfect inter-rater agreement or reliability. The study [28] reported a comparison of pairwise annotator performance for the i2b2 dataset with a κ coefficient.

Few studies reported another quality assessment test called K-Fold cross-validation to assess the model performance on unknown data. In this method, complete training data is distributed into several parts known as folds, and several iterations are run. In every iteration, initial fold is considered as testing data and the other leftover folds as training data. This process is continued until the last fold is considered as the testing data, and the score of each iteration is noted down. K-fold cross-validation method significantly reduces bias and variance as most of the data points are used in validation and training set at least once. Thus, the goal is to identify the best-suited data samples and improve data quality using K-fold cross-validation. Generally, for testing the data quality, the value $K = 5$ or 10 is usually preferred, which allows for choosing better samples [21], [36], [71], [72].

Table 10. summarizes the validation methods for automated information extraction in existing literature.

D. RQ4-AI APPROACHES USED FOR UNSTRUCTURED DOCUMENT PROCESSING

This section will provide a detailed survey on AI-based approaches available for automatic information extraction from unstructured documents. We will cover OCR, RPA and NER as the three approaches/techniques. Text extraction is the main stage in automating document image processing [22], [45], [76], [77]. The document images can be compressed or uncompressed, grayscale or color and the text in the images can be editable or non-editable [72], [75], [78].

A range of information extraction techniques are proposed for particular applications, containing metadata extraction from scientific journals [51], legal contract entity extraction [44], [46], receipt entity extraction [75], and clinical text extraction [28], [62]. It is quite challenging to design a general-purpose text information extraction system as there are a lot of variations in a document image. There may be complex layout images [24], [79], or images consisting of numerous variations in font style, font size, text color, text orientation, and text alignment [80], [81]. All such variations pose a great challenge to the problem of automatic text information extraction [47].

We now discuss the relevant literature that have addressed automatic text information extraction using OCR, RPA, and NER.

TABLE 10. Summary of validation methods for automated information extraction in existing literature.

Validation Method	Reference	Purpose	Value
Chi-square	[73], [74]	To measure the dependence between a feature and a class label and reduce number of features by removing irrelevant features,keeping those with highest score.	NA
Cohen's Kappa	[28]	Inter-annotator reliability by 4 annotators.	0.87
	[62]	Inter-rater agreement by 7 annotators	For well specified phenotype very high agreement :0.95, For other phenotypes: lower agreement 0.71
K-Fold cross-validation	[21]	Selecting most relevant data samples.	K=10
	[36]	Selecting most relevant data samples.	K=5 and K=10
	[71]	Selecting most relevant data samples.	K=10
	[72]	Selecting most relevant data samples.	K=5

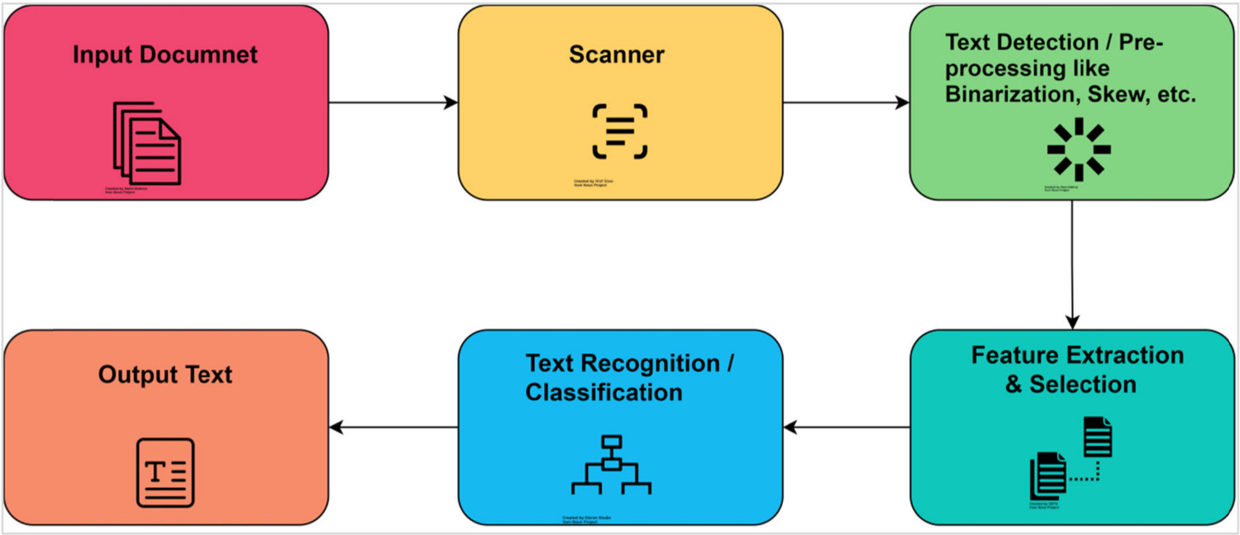

Input Documnet
Scanner
Text Detection / Pre-processing like Binarization, Skew, etc.
Feature Extraction & Selection
Text Recognition / Classification
Output Text

FIGURE 15. Steps in OCR.

1) OPTICAL CHARACTER RECOGNITION (OCR)

Manual data digitization is always laborious, time-taking and erroneous task. Utilizing OCR, organizations can digitize paper documents. It minimizes the need for human involvement in the less-significant task and increases the data reliability [24]. Organizations utilize this extracted text for analysis, processing, and editing. OCR involves two main stages. The first stage is text detection/localization in which the textual part within the image is located. This text localization within the image is essential for the second stage of OCR, text recognition, in which the text is extracted from the image. OCR can be classified as a handwriting character recognition system and printed character recognition system [82]. OCR for handwriting recognition is a complex problem because of the different writing styles and strokes of the letters used by the user. Discussion on OCR for handwritten recognition is out of the scope of this SLR.

We will now discuss the literature studied based on common steps in OCR as shown in Figure 15.

a: TEXT DETECTION OR LOCALIZATION TECHNIQUES

Text detection methods are necessary to identify or locate the text within the complete image and draw a bounding box over the portion or area of the image, consisting of textual contents. The text detection techniques are classified into conventional text detection methods and text detection using Deep Learning methods [83].

CONVENTIONAL TEXT DETECTION METHODS

Pre-processing is an essential step in text detection. Many researchers focused on pre-processing stages, like image resizing [83], blurring [72], thresholding [7], morphological operations [84], using OpenCV in conventional text detection methods. Pre-processing algorithms applied on the scanned document image depend on the several aspects such as image quality [18], scanned image resolution [1], [61], image skewness [18], [65], different format, and multiple layouts of the images and text [24], [78], [81].

Typical pre-processing includes the following stages:

- **Binarization:** In binarization, the grayscale images are transformed into binary images. It is necessary to recognize the objects of interest from the remaining part of the image. This text localization from the background is a prerequisite to the successive operations such as segmentation and labeling [18]. The easiest method is to compute a threshold value and change all pixels to white, that are exceeding that threshold value and the other pixels to black. OpenCV offers binarization via adaptive thresholding [7], simple thresholding [7], and Otsu's binarization [85].
- **Noise removal:** Scanned documents frequently consist of noise caused by the printer, scanner, and print quality [20]. For noise removal, the frequently used method is to process the image through a low-pass filter and use it for future processing [7]. OpenCV-Python can be used to get rid of such noise, such as salt & pepper noise [7] and Gaussian noise [65], [86].
- **Skew angle detection and correction:** When a document is scanned, either automatically or by a person scanning a document, a slight tilt (skew) to a document is obvious [35]. Given an image containing a tilted block of text at an unknown angle, the process of correcting text skew angle involves [19]:
  - Detecting the text block within the image.
  - Computing the angle of the tilted text.
  - Rotating or tilting the image to correct the skew.

Projection profile analysis [65], Hough transforms [17], [18], [65], and morphological transforms [20] are few methods for skew angle detection and correction mentioned in the literature.

- **Line-level, word-level, and character-level segmentation:** Segmentation divides the entire image into sub-images to process them further. The most popular techniques used for image segmentation are: X-Y-tree decomposition [20], connected component labeling [87], Hough transforms [88], and histogram projection techniques [7], [89].
- **Thinning:** Thinning aims to decrease the image parameters to its minimum necessary information, to simplify further processing or analysis and image recognition [19], [20]. It allows easier successive detection of relevant features. We found the most common thinning algorithm, the classical Hilditch algorithm, and its variations in a research study [17].

Conventional text detection techniques provide better performance; however, they still suffer from the following challenges:

- Conventional text detection techniques face a challenge to deal with the text images with complex image backgrounds, externally added noise, variations in lightning, diverse fonts, and geometrical misrepresentations or distortions.
- Conventional text detection techniques face a challenge to deal with unstructured textual content at arbitrary locations in a natural scene.
- Conventional text detection techniques cannot deal with unstructured text having complex layouts.

Some of these challenges are addressed using Deep Learning models. We will discuss Deep Learning models in next section.

TEXT DETECTION USING DEEP LEARNING

As discussed, in text detection, the text to be detected can be located in any image region. So, to localize the text, a bounding box is created around it. Various approaches for text detection are used once we localize the text with a bounding box. One such technique is the sliding window. In this approach, a window of appropriate size, say n x m, is selected to search the desired text over the target image. Few studies [18], [46], [67], [84] discusses a technique of creating a bounding box around the text with the sliding window. A sliding window slides over the image for text detection in that specific window. A sliding window then trains a CNN over every part of the input image. Different window sizes can be tried, so as not to miss the text portions or areas with different sizes. The disadvantage of sliding windows detection is its computational cost. Changing sliding window size and forming several square regions in the image, and running independently through a convolution network is computationally expensive. Efficient and Accurate Scene Text Detector (EAST) [32], [75], [83] is a Deep Learning (DL) model for detecting text from natural scenery images. It can discover horizontal and rotated bounding boxes. For a word or text line prediction the model containing a Fully Convolutional Network and a non-maximum suppression stage is used. CNN is used to extract features from the proposed image regions and outputs the bounding box and class labels. A non-maximum suppression stage is the last step of the object detection algorithms which selects the most appropriate bounding box for the object. It attains good accuracy in text detection. The resulting localized bounding text boxes are then given to OCR, for text extraction. The approach is slow since it requires a CNN-based feature extraction for each image area. In [79], a Deep Neural Network called ARU-Net is proposed to handle complex layouts and rounded and randomly oriented or tilted text lines of historical documents. Segmentation errors and false positives are the limitations of the proposed ARU-Net. We observed that few researchers have also developed their text detector using a customized text detector model. The TensorFlow Object API can be used to create a text detector [24]. TensorFlow is an ''open-source framework'' used to develop DL models for object detection tasks.

Although Deep Learning models for text detection provide improved performance, they still suffer from the following challenges-

- Deep Neural Network models always need a considerably large amount of annotated training data.
- Deep Neural Networks perform fine on standard datasets but can show poor results on real-world images outside the training data.

b: TEXT RECOGNITION

The next step is text recognition. In text recognition, the text characters are converted into various character encoding formats such as ASCII or Unicode. It can be performed in two steps (a) feature extraction and selection (b) classification

FEATURE EXTRACTION AND SELECTION

Feature extraction is learning and deriving the feature set, that accurately and distinctly describes the shape of a given character [17], [59]. Feature selection algorithms select the best feature subset from the input feature set [21]. Depending on the application to be developed, there are many feature extraction methods mentioned in the literature.

- **Template matching and correlation techniques:** A template or a prototype is considered as a representative of a particular character or object. Template matching is widely used method for extracting feature [38], [47], [72]. The template matching method uses individual image pixels as features. An input character image is matched with various templates or prototypes of each character class for accomplishing character classification. The closest matched template is assigned to that character. Template matching is used in many commercial OCR engines. However, it does not work well with noise and style variations. It fails to handle the rotated characters.
- **Structural approach:** In structural approaches, features that define the geometrical property of a character are extracted [76]. Character strokes, horizontal lines, vertical lines, endpoints, intersections between lines, and loops define the structural property of a character. Compared to other feature extraction and selection approaches, the features provided by the structural approaches are highly tolerant to noise and style variations. Structural methods use geometrical properties of a character and a classifier with some decision-rules to classify characters [20]. The structural features are less tolerant to character rotation and translation [22].
- **Statistical approach:** A character is denoted as a numeric feature vector in the statistical approach [24], [59]. It extracts the quantitative features like the total number of horizontal segments, vertical and diagonal segments, which are then passed to a classifier to classify the character. Different samples of a character are used for gathering statistics. The purpose is to provide all the shape variations of a character to the system. The character recognition algorithm uses this information, for classifying an unseen character. It is difficult to discriminate between the shapes of character due to the quantitative nature of the statistical approach.

CLASSIFICATION

The second phase in text recognition after feature selection is classification in which each character is identified and assigned to the correct character class. In simple terms, this process determines, what the character is. For example, suppose the structural approach gives the features consisting of one horizontal and one vertical segment or strokes. In that case, it might be either the character "L" or a "T," so to distinguish the shape of a character, the relationship between the two strokes is used. Classification, then assigns them a particular class. It identifies them either as "L" or a "T."

Various classification approaches used in the literature are discussed as follows:

- **Matching:** Matching consists of the groups of approaches based on calculating similarity distance. The distance between the feature vector and each class is calculated. A feature vector represents numerical feature vector describing the extracted character or object to be classified. The Euclidean distance is a commonly used matching technique because it is a minimum distance classifier [20].
- **Template-matching and correlation techniques:** The complete character acts as input to the classifier in the correlation approach. Since character features are not extracted, it is a template-matching too. In this technique, the distance between the input character image and a template is computed [13], [72].
- **Neural Networks:** Few studies focus on using Neural Networks (NN) to recognize characters [16], [72]. Recurrent Neural Network (RNN) and CNN are used for character classification and recognition for almost all languages [22], [90]. The disadvantage of NN in OCR is their limited prediction and generalization capabilities. However, the advantage of the NN is its adaptive nature.
- **Probabilistic approaches:** Statistical classification aims to apply a probability-based classification approach to text identification. Here, the aim is to use an optimal classification scheme [16]. One such optimal classifier that minimizes the total average loss is the Bayes' classifier [13]. Suppose a new character is given, which is represented by its numerical feature vector. The probability that the new character belongs to class 's' is calculated for all classes s = 1... N. The character is then allocated to the particular category or class with the highest probability.
- **Support Vector Machine:** It is mostly used in character classification problems in OCR. The kernel performs feature vectors mapping into higher dimensional feature space in SVM. A hyperplane is calculated, which linearly separates the classes by maximum margin. Kernel aims to transform the input data in the required output format by using various mathematical functions. SVM

TABLE 11. OCR engines used in existing research studies.

Category of OCR Engines	Name of OCR Engines
Open-source OCR engine	• Tesseract OCR [78] • Gamera Framework [59]
Commercial OCR	• ABBYY Finereader PDF [24] • Google Vision OCR [91] • Omnipage [76]

is considered the most robust and popular classification approach used in OCR for character recognition and classification [16], [20], [72].

c: THE OCR ENGINE

OCR engine is the software used for the text extraction from any image. They are either free /open source or commercial proprietary solutions. Each OCR engine comes with its strengths and weaknesses. Table 11. shows the existing literature studies that use an OCR engine for text extraction.

d: OCR ACCURACY

OCR accuracy is usually measured on character-level and word-level, that is, the rate at which a character/word is recognized correctly versus, the rate at which a character/word is recognized incorrectly. In the literature, we observed few methods to measure this character-level and word-level OCR accuracy [16], [24], [57], [59], [72], [91]. The most commonly used character-level accuracy is the edit distance. Levenshtein distance is the edit distance between two character strings [72], [91]. The term edit distance refers to the distance in which insertions and deletions has equal cost while substitution operation has twice the cost of an insertion. Given two strings n1 and n2, it is the minimum edit operations required to change string n1 to n2. So, if the edit distance is low, then the OCR engine has higher accuracy and vice-versa. In [57], [59], the Tesseract OCR engine is used to demonstrate the OCR accuracy on different datasets. OCR accuracy ranges from 90 to 95%, meaning that 10 or 5 out of 100 characters extracted by the Tesseract OCR engine are uncertain.

Another method to measure OCR accuracy is the confidence score. The confidence score is calculated based on OCR character-level and word-level accuracy scores combining with other existing information. This additional information can be the data type of extracted text, for example, numeric data type or letters as character data type or the text format. For example, the phone number format is different from the credit card number format [91]. OCR engines themselves compute the confidence score of the extracted text. Many OCR engines do not calculate consistent confidence scores, as they are incapable of taking additional information beyond character or word-level accuracy scores [91].

Another approach is based on selecting typical representative examples from the OCR engine extracted text and manual proof-reading the OCR output to correct the errors. It is a time-taking, error-prone, and tedious task in case of huge dataset [91].

We conclude from the surveyed approaches that the Levenshtein distance is the most simple and popularly used method for OCR accuracy. The first essential step is to ensure good quality source images are inputted into the OCR engine to improve the OCR accuracy. Thus pre-processing an image and layout analysis plays a crucial role in OCR accuracy.

e: BENEFITS AND LIMITATIONS OF OCR

BENEFITS

1. The organizations achieve higher productivity with OCR by simplifying the data retrieval process. OCR reduces the manual time and effort required to put in for extracting relevant data [2].
2. OCR helps the organizations to cut down the cost of hiring professionals to carry out data extraction [22].
3. Manual data entry is error-prone. OCR results in reduced errors resulting in an efficient data entry [6].
4. Scanned documents are always needed to be edited most of the time. OCR converts scanned document data to formats such as text, word, which can be easily edited [85].

LIMITATIONS

1. OCR can digitize text documents, making them machine-readable and editable. But, it cannot understand or interpret data [2].
2. OCR may not convert characters with very large or very small font sizes [24], [47].
3. OCR cannot process text inconsistency or variations in the layout of a document. It is a template-based method. Templates cannot handle probable complications like a printed and handwritten text combination, data within a table structure, and variations in text layout and formats [55], [56], [58].
4. OCR cannot extract non-textual characters or glyphs from the documents [20], [24].
5. OCR accuracy also depends on good quality source images [91].
6. OCR alone is not efficient for end-to-end automation in organizations. In combination with RPA and AI, OCR is a better solution for the unstructured document processing [2].

2) ROBOTICS PROCESS AUTOMATION (RPA)

RPA is the automation technology used for software tools or bots that automate human tasks, which are manual, rule-based, or repetitive [92], [93]. The word ‘Robot’ in ’RPA’ is

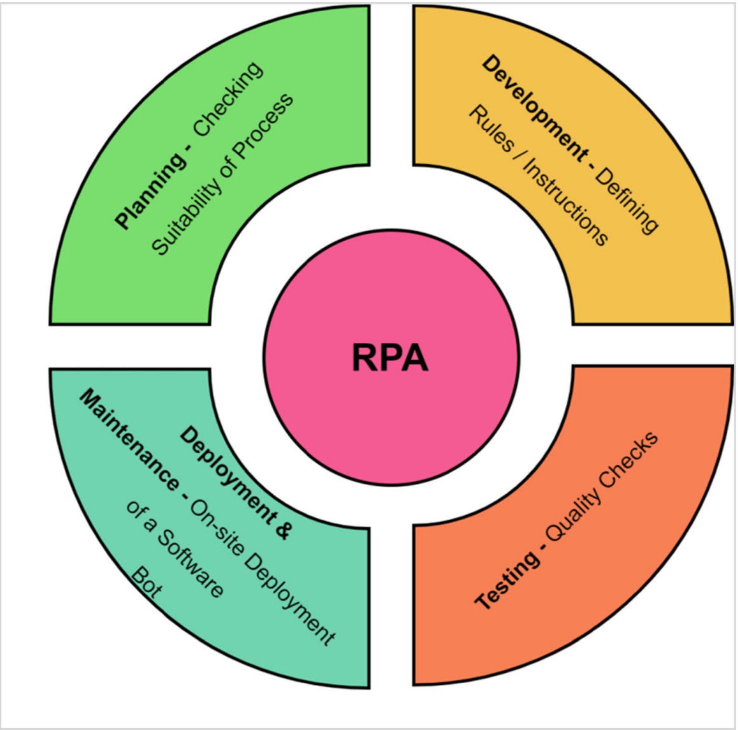

Planning - Checking Suitability of Process
Development - Defining Rules / Instructions
RPA
Deployment & Maintenance - On-site Deployment of a Software Bot
Testing - Quality Checks

FIGURE 16. RPA implementation steps.

not a physical robot but a virtual system, that automates repetitive manual tasks or business process tasks. It performs such tasks quicker than a human, with no mistakes. It can interact with websites, user portals to login into applications, enter data, open emails, attachments, calculate and complete tasks, and then log out. It is known as one of the most pioneering technology. It is growing very fast, as organizations try to do things easier and faster with software bots [8].

a: WHY RPA?

RPA aims to automate business processes to improve the efficiency by reducing the costs and efforts, humans spend on repetitive tasks. An example of such repetitive tasks is logging in to applications, typing, copying, extracting, and moving files or documents from one system to another. A software bot can do such structured and manual tasks [25]. RPA can help to automate digital business processes with an accurate decision making [14], [94]. It can streamline repetitive and rule-based business processes [25]. It enables systems to make intelligent decisions, with the help of RPA software robots. Recent literature studies discussed the advantages of the RPA implementation in several application domains regarding increased productivity, reduced operational costs, improved service quality, and error reduction [93].

b: BPM VERSUS RPA

The goal of Business Process Management (BPM) is to re-engineer process workflows, to eliminate bottlenecks, increase productivity, connect different systems of organizations. After process reengineering, BPM requires to build a new application for interacting with other applications using Application Programming Interface (APIs). In contrast, RPA aims to automate existing processes, that are already well-defined, standardized, and usually executed by the human workforce. RPA provides Graphical User Interface (GUI) to integrate with the other applications. In RPA, a new application need not be developed, to integrate with existing systems [2], [25].

c: RPA IMPLEMENTATION STEPS [25], [27], [95]

This section discusses the process identification criteria and general steps for RPA implementation. The processes that meet these requirements could be automated and increases the operational efficiency of any organization. Figure 16. shows the RPA implementation steps, which are discussed as follows.

- **Planning:** In the planning phase, organizations identify the candidate processes, that they want to automate. Along with identifying the correct process, the organization should check the suitability of the existing system for RPA implementation.

The candidate process should satisfy specific criteria for RPA automation:

1. The process should be high volume, manual & repetitive. For example, thousands of documents are processed every month, involving much human workforce in that business process.
2. The process should be rule-based, that is, a definite set of rules are needed to complete the process.

3. The process must be standardized, that is, a process must have specifications that are standard and fixed most of the time. For example, it must be known what data fields should be extracted from a document.
4. The input data must be clear, unambiguous, understandable, and in electronic format.

- **Development:** Developing automation workflows as per requirements.
- **Testing phase:** Quality check on RPA to identify and correct defects.
- **Deployment and maintenance:** In this phase, deployment and maintenance of the software bots are performed to detect any exception or flaw for immediate resolution.

d: COGNITIVE RPA

This section discusses various case-studies or uses of cognitive RPA mentioned in the literature:

- **Cognitive RPA meaning**

RPA is rule-based automation technology. Without intelligent software additions, RPA cannot process the unstructured documents such as invoices, contracts. RPA is called cognitive RPA, when used in combination with AI technologies such as OCR, NLP, and ML, to improve the process workflow for end-to-end automation [10], [27].

- **Cognitive RPA use-cases**

RPA technique proposed in the study [10] is based on DL model, and it can detect objects in real-time, classify them with high accuracy, and take actions dynamically. It shows how RPA mimics human actions while executing various tasks within a process, such as clicking on the help or file menu button. It indicates that, it is possible to automate any computer task or user activities with the RPA implementation. The CNN is used as an underlying DL model trained with numerous interfaces and menus to classify given software interfaces in real-time.

The document processing is important to the entire operational workflow in many businesses. For example, in healthcare applications, member enrolment form processing, Electronic Health Records (EHR) management, claims processing, and other activities require analysis of information. RPA is used to automate these tasks [40].

The study [27] discusses automating the business process from a debt collector company. The company receives more than 1,50,000 documents each month for two categories, court and bailiff. Before RPA implementation, each document gets scanned manually and assigns the main category to it: either 'Bailiff document' or 'Court document.' The process is automated, by using OCR and NLP along with RPA implementation.

Two case studies on the RPA implementation using unstructured interview documents are found in the study [41]. The multinational company is using RPA, to automate the customer onboarding process. Another technology and consulting company is using RPA to automate the responses of interviews.

e: USE OF RPA [8], [26], [93]

1. **Imitates user activities:** RPA can automate the execution of the repetitive business process.

2. **High-volume repetitive tasks:** RPA can automate the high-volume repetitive task quickly and reduce errors. For example, copying or moving data from one system to another, data entry operations, and extracting specific fields from the documents.

3. **Multiple Tasks:** RPA can automate multiple and complex business processes across several systems. For example, processing transaction updates, manipulating data, and sending updated reports.

4. **Automatic generation of reports:** RPA can automatically extract the data fields required to generate accurate, useful, and timely reports.

f: RPA TOOLS

Few research studies discuss RPA tools and its vendors [14], [25]–[27]. Organizations can use RPA tools from different RPA vendors. RPA Tools are widely used for the configuration of automation tasks. These tools are crucial for the automation of repetitive back-office processes. Several RPA tools are available in the market, and choosing the right tool could be a challenge.

The following parameters are considered to select the right RPA tool:

- The RPA tool should able to read and write business data into multiple systems.
- It should be easy to configure on rule-based or knowledge-based processes.
- It must work across multiple applications or systems.
- It should have built-in AI support to repeat or mimic the activities of the user.

RPA Vendors: 'Blue Prism', 'UiPath', 'Automation Anywhere', 'Verint', 'WorkFusion', 'Kofax', 'Weka', 'Cloudera' are few mentioned leading RPA vendors [26], [27].

g: BENEFITS AND LIMITATIONS OF RPA

BENEFITS

1. RPA can automate a large number of processes smoothly [26].
2. RPA reduces the operational cost significantly as it can handle repetitive tasks and saves valuable time and other resources [14].
3. Prior programming knowledge is not necessary to set-up and implement RPA. Thus, any non-technical person, without any programming knowledge can operate the RPA interface to automate the process [93].
4. RPA supports almost all the regular business processes with error-free automation [26].
5. RPA significantly reduces human intervention [25].
6. Software bots or tools do not get exhausted. Software bots are scalable [14].

LIMITATIONS

1. RPA is suitable for processes that include rule-based tasks. The business process working details and logic

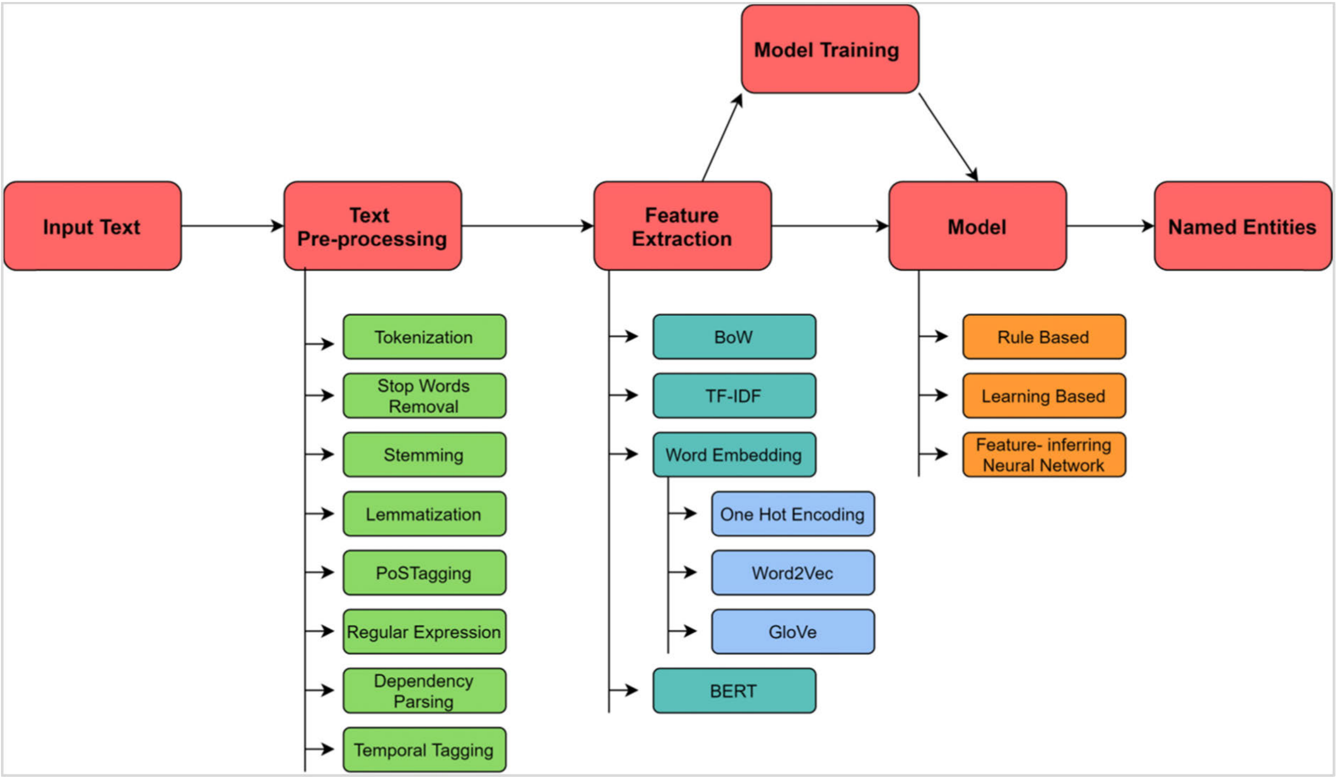


FIGURE 17. NER workflow.

behind its functionality need to be expressed in terms of instructions or rules. RPA requires defining specific rules for every case, which must be clearly defined and unambiguous [25].

2. RPA does not work well for multiple layouts or formats of the unstructured documents, with the text fields placed in different locations or places inside a document. For example, if a 'software bot' need to read an invoice, all different supplier invoices should follow the same layout format, with the same type of fields. Although software robots can be trained for exception handling, to read different locations or fields inside a document, they fail to read multiple formats [41].
3. RPA software bot needs to be reconfigured even for small changes made in the automation application [14].

3) NAMED ENTITY RECOGNITION (NER)

NER automatically scans the unstructured text to locate the "named entities" like a name (first name, last name), location (such as countries, cities), organization, date, and invoice numbers in the text [30]. For example, "Siddharth Properties" is identified as "entity" from other text and assigned to the category "organization," and "Nilesh" is identified as "entity" and assigned to the category "person." "Named Entity Recognition and Classification" is another term used to refer to NER [29]. Business document NER is challenging because such documents may contain non-standard phrases describing entities such as invoice_Num or invoice_No, for representing the same entity [22].

a: NER WORKFLOW

A NER system can identify entity elements from the text and decide its class. General entities like name, organization, date, and location can be identified and classified using Stanford NER and Spacy [12]. For identifying and classifying domain-specific entities, the NER model is trained using custom/self-built training data. Creating a custom/self-built dataset is always a tedious task, requiring lots of human effort and time for annotations.

The main steps in NER workflow include, text preparation and model training [29]. Text preparation includes, text pre-processing and feature extraction. The Figure 17. shows the NER workflow.

TEXT PREPARATION

Text preparation consists of two sub-phases: A) Text pre-processing and B) Feature extraction.

A) Text Pre-Processing

Text pre-processing involves cleaning or preparing the given data for processing, such as removing stop words and stemming. It also consists of breaking a sentence into tokens. Data normalization is needed to reduce the ambiguity, which may later affect the feature extraction step. Data normalization consists of tasks such as stemming, lemmatization, upper or lower casing, and stop words removal.

A text document is given input to NLP tools like "NLTK" or "spaCy" to convert the character sequences to normalized tokens. Following text-preprocessing techniques are found in the literature for such conversion:

- **Tokenization:** Separating a sentence into smaller elements called "tokens." Tokens are words or subwords, or characters [12]. For example, given the sentence "This is a text," the tokenization breaks the sentence into smaller components called tokens as shown- ["This," "is," "a," "text"]
- **Stop-words removal:** A stop word is a usually used auxiliary verb, conjunction, and articles (such as "the," "a," "an," "in"). They occupy memory in the database and consume processing time. So, stop-words need to be removed from sentences [96].
- **Stemming:** Reducing a word to their word stem called as a lemma, or base, or root form [63]. It can be merely stripping of recognized prefixes and suffixes. For example, take is the stem word for taking. Removing 'ing' is an example of stemming.
- **Lemmatization:** It is a dictionary-based approach to determine the lemma or base or root form. It is recommended over stemming, when the meaning of the word is essential for analysis [11], [71], [97]–[99]. For example, the root form of "jumping" is "jump," and the root form of "are" is "be."
- **Part of Speech Tagging (POS):** Allocating "tags" or "parts of speech" to every token, such as a "noun," "verb," and "preposition" in a sentence [63]. For example, "**NN"** is the tag for a singular noun. For the tokens ['Hello', 'world'], the POS tagging output is [('Hello', 'NN'), ('world', 'NN')].
- **Regular Expressions:** Regular expressions are the patterns or set of characters in the form of an instruction given to a function, to find a substring or match the strings or replace a set of strings. It uses particular notations. For a pattern, the character (-) specifies a range, and the character ('?') represents zero or more occurrences of a particular character in a string [100].
- **Dependency Parsing:** It finds the relationships between two words in a sentence represented by the various tags [101]. For example, given the sentence to find the dependency parsing- 'I prefer the morning flight through Mumbai,' it states that 'the' is used as 'determiner' for 'flight.'
- **Temporal Tagging:** It is the task of finding phrases with temporal meaning or temporal expression [22]. For a sample sentence, "I may go to the college in the next two weeks," the temporal tag-detected is "the next two weeks."

B) Feature Extraction

NLP model cannot work on the raw text data directly. So, feature extraction methods are required to convert text into a numerical representation of features or a matrix (or vector) of features [66], [96], [102]. It maps words into numerical vector space, which is considered a richer representation of text input in NLP.

Various methods found in the literature for feature extraction are discussed as follows:

- **Bag of Words (BoW):** It calculates the word occurrence/frequency in a given document. In simple terms, it counts the word appearance in a given document [61], [98], [104], [105]. The word counts are used as a feature for training a classifier [61]. For example, in [21], a BoW is used for business invoice recognition and to capture layout and textual properties for interested fields.
- **Term Frequency-Inverse Document Frequency (TF-IDF):** It calculates the word occurrence/frequency referred to as "Term Frequency" inside the entire document, against the word occurrence/frequency count inside the document corpus [44]. In TF-IDF, weights are assigned to words. It is usually used to get the relevance of words. Common words such as "and" or "the" are frequently used in all the documents. Those words count must be skipped. It is the "Inverse Document Frequency" part. If the word appears in more documents, the word is counted as less important, as a signal to differentiate any given document. The distinctive words are then given input to the Neural Network as features to conclude the "topic covered" by the document [105].
- **Word embedding:** BoW and TF-IDF approaches do not capture the meaning or relation amongst words from vectors. Word embeddings can capture semantic, syntactic relationships between words and the context of words in a document.

We found four word embedding approaches in the literature, which are explained below:

1. **One Hot Encoding:** One-Hot-Encoding vector representation is the most simple and fundamental word embedding technique. Categorical/nominal variables are represented as binary vectors using one-hot-encoding. For that, the categorical values are represented with numeral values first. Many AI-based algorithms do not work with categorical/nominal data directly. So, categories must be transformed into integers for both input and output variables, that are categorical. For a size 'S' word vocabulary, every word is given a binary vector of size 'S,' where all vectors are 'zero' apart from one related to the index of the word. Typically, the word index is found by positioning all words with a rank, where the rank relates to the index. For example, consider a label sequence ['blue,' 'yellow']. Assign integer value 0 to 'blue' and the integer value 1 to 'yellow.' The length of the binary vector will be 'two' for the two possible integer values. The 'blue' label encoded as 'zero.' It is denoted by a binary vector [1, 0], so the "zeroth index" is written with a value '1'. The 'yellow' label encoded as 'one.' It is denoted by a binary vector [0, 1], so the "first index" is written with a value '1'. The main concern of this type of encoding is the word vector size. For a bigger corpus, word vectors are extremely big-size and exceptionally scattered [79], [106]–[110].

2. **Word2Vec:** Word2Vec is a standard and popular technique to create word embeddings. Word2vec is a two-layer shallow Neural Network used for processing the text by word vectorization. It takes a text corpus as input, and outputs the feature vector representation of words in that text corpus. It converts text into a numerical vector representation that Deep Neural Networks can recognize. It groups the vectors of similar words. It identifies similarities between the words mathematically. Word2Vec architecture has two algorithms, that is, Continuous Bag-of-Words (CBoW) and Continuous Skip-Gram. CBoW works well with large datasets. It is used to get representation for frequent words than rarer words. Whereas, Skip-Gram works well with small datasets. It identifies rarer words better. Based on the application requirements, any one of the approaches is used [29], [34], [62].
3. **GloVe:** GloVe (Global Vectors) is a widely used word embedding method used to get its vector representation. It learns word vectors by performing dimensionality reduction on a co-occurrence counts matrix. GloVe builds a co-occurrence matrix for the complete corpus first, then factorizes it to create matrices for *word vectors* and *context vectors*. Compared to the Word2Vec approach, parallel implementation is possible in GloVe. It implies that it is useful to train over more data [29], [34], [111]–[113].
4. **BERT:** Bidirectional Encoder Representations for Transformers (BERT) [12], [29], [34], [111], [112], is nowadays the latest word embedding approach, that is effectively used in numerous biomedical and other text mining tasks. BERT learns the text representation from both the directions to better understand the context and the relationship. It encodes the word context by having information about the previous and next word in the feature vector representation. BERT gives much-improved NLP task results, as it can understand the word context and work well on the unlabelled corpus.

2) NER MODELS

As we have discussed, the extracted features are passed through a NER model, that will classify different words and phrases into specific categories. The unstructured text is rich with information, but finding relevant data from it is always challenging. NER is the best choice to categorize the data in a structured manner. It pulls out entities from such data and assigns suitable categories to them.

The study [12] categorized NER techniques into three main methods/approaches 1. A rule or knowledge-based approaches, 2. Learning-based approaches, and 3. Feature-inferring Neural Network approaches.

1. **Rule or knowledge-based approaches:** Rule-based approaches provide benefit because they do not require large annotated or labeled data for training, but they depend on lexical resources [12]. A lexical resource consists of the numerous dictionaries consisting of various stop words, commonly used words, and homophones. For example, defining regular expressions using "pattern" or "set of characters" and deriving context-free grammar are rule-based approaches in NLP. These approaches generally show low precision, high recall. It means they work well for specific use cases, but not for generalized-task [22], [34].
2. **Learning-based approaches:** Few studies have discussed the learning-based approaches [31], [34], [44]. These methods are used to substitute the human-defined rules, that are necessary for a rule-based category. Algorithms learn and understand the language from the annotated corpus or training set, to produce its own rules and classifiers. It is possible through the use of statistical methods. Learning-based methods are classified into three types of classes: "**supervised learning", "semi-supervised learning", and "unsupervised learning"**. **A supervised learning** algorithm aims to classify identified named entities to be correctly classified into their categories. Supervised learning is used, when the model is trained by the data that is very well labeled. It means certain data is already labeled with the correct solution. Few supervised learning-based methods reported in the literature are Hidden Markov Model (HMM) [44], [114], SVM [34], [44], Decision Trees [83], and Naive Bayesian methods [29], [115], Conditional Random Fields (CRF) [46], [113]. **Semi-supervised learning** algorithms iteratively apply a supervised learning algorithm with both labeled and unlabelled data. The classifier learns to classify unlabelled data based on the knowledge of the labeled data. A semi-supervised model aims to classify some of the unlabelled data using the labeled data [44]. **Unsupervised learning** algorithms discover patterns in the data on their own by learning from data. In that sense, they are automatic. However, they essentially need a minimum amount of training data [12]. In the unsupervised learning, the model works on its own, to learn information or knowledge. It mostly deals with the unlabelled data. The unsupervised learning algorithms use input documents to find structure or pattern by observing the association between the inputs documents. However, very few studies used this type of learning [37], [43], [51], [111].
3. **Feature-inferring Neural Network approaches:** The third category, feature-inferring Neural Network approaches, differs from other two approaches in utilizing and extracting features from DL models. The development in feature-inferring Neural Network approaches is significantly helping to analyze the unstructured documents from the past few years. DL models are generally trained end-to-end, manual feature extraction is not required. In a Deep Neural Network, lower layers learns the feature representation on their own, and the

higher layers can work as a classifier. RNNs and CNNs are specific and popular Neural Networks used in NLP. We now discuss how RNN, CNN, and their variants are used in the NER model literature.

- **Recurrent Neural Networks (RNNs):** RNNs are the most well-known Deep Neural Network learning model to solve the NER task. In RNNs, all weights/parameters are used again for each iteration stage. The next hidden state output is updated, depending on the previous hidden state inputs as well as the current input. The hidden state acts as the memory of the neural network, storing the most important information from the previous inputs. RNN is used for Biomedical Named Entity Recognition (BioNER) systems in the study [29]. It extracts the relation between the biological entities to recognize the interactions among proteins and drugs or genes and diseases. In [83], combination of CNN and RNN, called as Convolutional Recurrent Neural Network (CRNN) is proposed for extracting the textual information from the images of medical laboratory reports, which might help physicians to check the details of the patient.
- **LSTM & GRU:** LSTM and GRU are the types of RNN. Simple RNNs have a very short memory. To solve the short memory problem, more complicated Neural Network architectures are used. The most famous ones are the LSTM as well as GRU. LSTM comprises Neural Networks and various memory blocks called cells. Through these cell states, the information flows. LSTMs can specifically remember or forget things with these cell states. The cells reserve information and the gates do the memory operations. The Gated Recurrent Units (GRU) have a slightly simpler, less complex architecture. It consists of a hidden layer and gates that control data flow. GRU has quite similar properties to LSTM. Both LSTM and GRU uses a gating mechanism to perform memory-related functions. The study [28] proposes LSTM and GRU to develop supervised learning Natural Language Processing (NLP) models to extract symptoms or diseases from the unstructured clinical notes dataset. We found few other studies also using LSTM and GRU for information extraction task [29], [50], [57], [68], [110].
- **Bi-LSTM & CRF:** A bidirectional LSTM is a type of RNN. It has two LSTMs consisting of a forward layer and a backward layer. The forward layer has a text sequence input to it. The backward layer processes the input in the reverse order. It starts processing the last word, then continuing to the next to the last word, and so on to the first word. The hidden states combine each token producing an intermediary representation sequence. Therefore, each intermediary representation of the information from the sequence, before and after the individual token is considered. For each iteration step, the network can process the complete document and infer the right label from that information. Conditional Random Field (CRF) models the text sequence. It assumes that features are dependent on each other. It is based on calculating and maximizing the conditional probability to predict the correct text sequence. CRF in NLP is used in NER for sequence prediction, where features depend on each other. We found few studies mentioned the advantage of Bi-LSTM and CRF for information extraction task. The study [35] focuses on the bidirectional LSTM–CRF model to achieve better clinical named entity recognition performance. In this study, multitask attention based BiLSTM-CRF model combined with context-based word representation is used. An attention-based neural network considers two sentences and transforms them into a matrix format, in which columns represent the words of one sentence and rows represent another sentence. After this, to identify the relevant context, it performs word matching. BiLSTM–CRF based Model used in this study, shows the improvement in recall value in the entity discovery task. The study [68] aims to improve the text classification accuracy by combining an attention-based Bi-LSTM, and CNN called a Hybrid model. It combines the features of LSTM and CNN along with an attention-based mechanism. The classification model trains on IMDB a movie review dataset.
- **Convolutional Neural Networks (CNN):** CNNs are neural networks used mainly to classify images and feature extraction to identify and recognize text lines on identity documents dataset MIDV500, MNIST, and custom dataset in the study [57]. The study [36] proposed invoice classification in three invoice classes: *machine-printed invoice, handwritten invoice, and receipts*. Alexnet, a Deep CNN, is used for feature extraction from invoices. It is used as a pre-trained model on the Imagenet dataset and later on the invoice dataset. Convolutional Networks can also perform text digitization using OCR. CNN and NLP are combined and used to extract information from business-oriented scanned documents such as invoices and purchase orders [36]. The study [47] used the R-CNN for locating objects in real-time. They are well-known for detecting objects as their object detection speed is relatively high. The study [34] used a CNN, that combines word embedding for named entity extraction from clinical notes. The study [68] used the convolution layer and pooling layers for information pre-processing before giving it to LSTM to process the input and construct better features set, to process long sentences. It helps to improve the LSTM efficiency and effectiveness. The study [10] proposed an RPA application for dynamic object detection from the software applications interface. Numerous interfaces and menus are given input to train a CNN for "dynamic object detection". CNN is also used here for real-time software interface classification. The tool CNN YOLO (You Only Look Once) is used in TensorFlow for detecting objects in real-time and feature extraction, allowing the decision-making process and performing real-time

action. The YOLO algorithm performs far better than any other algorithm, as its training and classification time is very less for real-time object detection and reaction. CNN YOLO is also used for object recognition and feature extraction in complex documents such as comics [87].

- **BERT:** Few studies [109], [116] reported the use of BERT for extracting features from the data. It is suitable for various types of language processing related task in NLP. The model is usually trained on an unannotated dataset for transfer-learning. Fine-tuning this pre-trained neural network model can be utilized for different NLP tasks like sentiment analysis, sentence classification, question answering, and few others. It is designed for pre-training the Deep Bidirectional Neural Networks from the unlabelled text in NLP. BERT uses a transformer, an attention-based mechanism that learns contextual relationships between the words in a text. The main advantage of using a transformer encoder is that, it reads the entire word sequence simultaneously, unlike unidirectional models, which read the input text serially either from left-to-right or from right-to-left order. Though BERT is a powerful NLP model, using it for NER without fine-tuning it on the NER dataset will not give better results. Fine-tuning BERT for NER performance improvement in financial and biomedical documents utilizing the combination of BERT and word embedding is discussed in the study [109]. Utilizing and fine-tuning the BERT model for achieving document-level relation extraction using DocRED-a large scale open-domain document-level relation extraction dataset shows improvement in F1 measure in [116].

E. RQ5-UNSTRUCTURED DOCUMENT PROCESSING TOOLS

A few studies reported commercial tools developed to fulfill the need for an automatic extraction of useful information/fields from the unstructured documents [13], [33], [51], [72], [89]. The organizations aimed to extract relevant and specific information from the unstructured documents such as invoices, orders, and credit notes. These are commercial tools or frameworks. Hence, the literature lacks the details of the steps involved in developing these tools or frameworks, the datasets they used, the techniques they applied, or the evaluation metrics they used.

One of the most impressive commercial tools, Cloudscan, is discussed in the study [90]. It is an invoice analysis system with a Graphical User Interface(GUI) with zero configuration and requires no upfront annotation. It takes a PDF file as the input, extracts the words and their positions, creates N-grams of words, and extracts the text features from N-grams generated. LSTM is used to classify each N-grams 32 fields of interest. The classified region of interest from LSTM is then given input to the post-processor, which filters out the N-grams, which does not fit with syntax with regular expression parsers. Finally, the results are exported to Universal Business Language (UBL). The output is a UBL invoice. This UBL invoice is added to the CloudScan data collection as a training dataset. The N-grams and the labels from the user validated UDL invoices are used to train the classifier. CloudScan usually works well, since it has a huge collection of the dataset, as mentioned above. The advantage of CloudScan is that it offers an easy user interface. However, LSTM and feature engineering used in CloudScan still has a scope for improvements. Word-level processing is a feature of CloudScan. CloudScan fails in image feature processing like the logos and watermark or any image background.

Another tool reported in the study [33] is Convolutional Universal Text Information Extractor (CUTIE). CUTIE learns from the input data provided to it for training. It requires significantly less human intervention. It uses DL techniques without defining rules for any specific type of document. CUTIE can process simultaneously on both semantic information and relative position coordinates (spatial distribution) of texts. It works on "gridded texts." The purpose of the grid is to generate a matrix-like or a tabular structure, where the text is a "feature" with semantic information. The grid also preserves the relative co-ordinate position relationship of text from the original scanned document. For the scanned document image, the gridded text is created, when OCR outputs the extracted texts with their relative position coordinates. The word embedding layer is used to get the semantic information from these gridded texts. The Scanned Receipts OCR and Information Extraction (SROIE) dataset provided in the conference titled The International Conference on Document Analysis and Recognition-2019 (ICDAR) and a custom dataset consisting of three categories of scanned document images are used for evaluation of CUTIE framework.

Apart from the above tools, few studies reported the task-specific tools for the information extraction. The review [13] summarized the clinical information extraction tools used in the various related studies. The authors highlighted few widely used tools for extracting clinical information in the medical area: cTAKES, MetaMap, and MedLEE.

Similarly, CERMINE [51] is an extensive, publicly available non-proprietary tool for extracting structured metadata from electronic (digital) scientific publications. CERMINE can process publication document layouts with complex variations quite well. A PDF format scientific publication is given as, input to CERMINE. The information extraction algorithm in CERMINE scans the whole document content and creates two types of results: The details of the scientific document in metadata and its bibliographical data. CERMINE gives metadata such as document heading, information of the writer, publication abstract, and other bibliographic information.

Intellix by DocuWare [33], [90] is a document processing tool for classification and information extraction. It needs an annotated template with related fields. A collection of templates must be created in a structured form, that is, references and affiliations and their metadata. It classifies the input document into certain classes such as *invoice, customer communication letters, or delivery*. It is a training-based

TABLE 12. Tools/frameworks developed in prior research studies.

Framework/Tool Name	References	Techniques Used	Accessibility to Framework/Tool	Features	Drawbacks
Convolutional Universal Text Information Extractor (CUTIE)	[33]	OCR and Convolutional Neural Networks (CNN).	Prototype	For scanned document image, CUTIE can process semantic as well as relative or absolute position (spatial) statistics of the texts simultaneously.	Faces difficulty in model inference when entity names vary greatly.
CloudScan	[90]	RNN-based classifier, bi-directional LSTM model.	Prototype	Extract necessary information fields with learning-based models without any template for invoice documents.	Faces a challenge in finding or learning the relationship between the words.
Intellix by DocuWare	[33], [90]	Machine Learning.	Commercial	Eliminates the need for manual document filing.	Rule-based invoice analysis systems.
CERMINE Tool	[51]	Optical Character Recognition, Support Vector Machine.	Open-access	Extracting structured metadata from references.	Can not handle scanned document information extraction.
DoCA (Document Classification and Analysis) framework	[72]	CNN,LSTM model.	Prototype	A tool to analyze and classify documents in different file types.	Template-matching is used

TABLE 13. Comparison of OCR, RPA, and AI-based approaches.

Parameter	Optical Character Recognition	Robotics Process Automation	Artificial Intelligence
Ability to handle input data type	Works well on structured and semi-structured good quality scanned images.	Works well on structured data and some semi-structured data (Spreadsheets, RFID tags, GPS data.)	Works well on all data types, including unstructured data (word files, emails, images.)
Approach	Rule or Template-based.	Rule-based.	Learning-based : learns from data collected.
Processing approach	-	Deterministic.	Probabilistic.
Technology	Text detection and recognition.	Software robots configured to perform the repetitive task and complete routine.	AI-based on Deep Learning, Machine Learning, Computer Vision, Text Analytics, NLP.
Automation scope	Allows converting diverse categories of documents, like scanned documents, PDFs, or images, into editable and searchable data.	Minimize manual, repetitive, and rule-based task.	Can automate tasks that require decision making and increase the scope of automation.
Benefits	Editable and searchable data, which can be further utilized for storing in databases.	Business benefits are quick and immediate, returns can be realized quickly and cost-effectively.	Scalable and flexible.
Output	The key tool for digitizing documents.	Completed task or process, exception or alerts.	Structured output for the advanced integration.

information extraction tool purely based on text and layout features.

DoCA (Document Classification and Analysis) [72] is a framework for the classification and analysis of diverse file types comprising of office documents such as text files, excel sheets, PowerPoint presentations, scanned PDFs, multimedia files like audio and video. It is a template-matching based framework. HAVELSAN dataset is used for studying the effectiveness and feasibility of the DoCA framework. Table 12. shows some tools used or developed by the researchers in the automatic information extraction from unstructured documents.

F. COMPARISION OF OCR, RPA, AND AI-BASED APPROACHES

Table 13. shows the comparison of the OCR, the RPA, and the AI approaches. This comparison is based on different parameters mentioned in the table. Studied approaches are highly complementary with each other and fall under the broader field of automated information extraction research. Depending on the need and application of the organization, these approaches can be combined to implement end-to-end process automation solutions.

It is inferred from the Table 13. that OCR is a template-based method used for text recognition for a very long time. RPA is used to automate rule-based tasks. Both OCR and RPA works well on structured data. AI-based techniques deals with the unstructured data making the hidden contents more useful. It can mimic human intelligence and can easily extract values from the unstructured data. AI-based techniques are proficient in ''understanding'' the text, classifying it accurately. Based on that classification, AI-based techniques also helps to create automated workflows without human intervention. With the combination of OCR, RPA and

TABLE 14. Summary of reviewed studies.

Technique	Purpose	Domain	Feature Extraction Methods	Features Used	Hyperparameters	Results
[68] Bi-LSTM, CNN hybrid	Text classification	Internet Movie Database (IMDB) movie review	Word2vec's skip-gram model	Sequence of words	Embedding size: 500 batch size: 128 dropout ratio : 0.2	F1-score-0.9018 Recall-0.9057 Precision-0.8975
[57] CNN , ANN	Identity documents image recognition	MNIST, MIDV-500 and Census 1961	CNN as character classifier	Character	-	OCR Accuracy: 95.40 %
[104] CNN	Handwritten Urdu character recognition	UNHD, EMILLE, DBAHCL, and CDB/Farsi	AlexNet, GoogleNet ResNet18	Character	-	Character recognition accuracy: 97.18%
[62] CNN	Text classification of clinical database	MIMIC-III database	CNN	Disease symptoms	-	F1-score: between 57 and 97 %
[114] CNN, RNN	Medical invoice recognition	3755 commonly used Chinese characters	Scale-invariant feature transform (SIFT)	Invoice words	Base learning rate: 0.01, iteration: 8000	Accuracy: 70 to 85%
[70] CNN, BoW/SVM classifier	Defining text's category and categorization	20Newsgroups dataset	TFIDF BoW features	Words	Mini-batch stochastic gradient descent l2-regularization	Accuracy: 80%
[117] R-FCNN	Object detection.	PASCAL VOC, MS COCO datasets	RoI computation	Image features	Learning rate - 0.001- for 90k iterations 0.0001- for the rest 30k iterations, mini-batch size - 8	mean Average Precision: 83.6%
[47] image processing techniques, and template matching	Automatically identify Chinese invoice information	20 invoice images	Geometric positional relationship,edge detection,counter extraction	Invoice number and billing date	-	Accuracy 95%
[21] Naive Bayes, Logistic Regression, and Support Vector Machines (SVM)	Invoice recognition using layout analysis	97 raw invoice images, Oracle Corporation	Bag-of-words approach	Invoice details	l2-regularization C = 10.72: Logistic Regression C = 0.58 :SVM.	Training error: 8.81% Testing error:13.99%
[61] CNN	Document image classification and retrieval	IIT-CDIP collection	BoW, CNN	Image features	Convolutional layers:5, fully-connected layers:3	Accuracy: 95% mean Average Precision: 0.799
[46] BLSTM, LSTM, BLSTM-CRF	Contract element extraction	3,500 contracts documents dataset labeled with 11 types of contract elements in English	Word2vec's skip-gram model	Contract and contracting parties	Output gates with dropout, binary cross-entropy loss, Adam optimizer	F1-score-BLSTM-LSTM-0.97 Recall-0.95 Precision-0.93

TABLE 14. ***(Continued.)*** **Summary of reviewed studies.**

[31] 2D-LSTM, R-CNN, R-FCN	Detection of text in document images and directly predict object coordinates	Maurdor- highly heterogeneous dataset	CNN as text classifier	Text	-	Recall: 35.1- 54.2, Precision: 38.4 – 61.4, F1- 36.7-56.3
[37] OCR, Concept Lattice	Text Information extraction with extracting the relevant text data from a collection of document images	The synthetic data set of 1000 invoices, i.e., 1000 real-word invoice images	ABBYY OCR, Tesseract OCR	Document images text	-	Tesseract OCR engine accuracy: 70%, p-value 8.799e-05
[36] Alexnet, OCR, Classification algorithms	Invoice recognition	Invoice documents	AlexNet, KNN, Random Forest, Naïve Bayes	Invoice words	-	Accuracy: 94%
[34] LSTM–CRF,NER	Information extraction from Medical records	I2B2/VA	RNN & LSTM	Words	-	F1: 88.78 Precision: 87.75 Recall: 88.46
[33] CloudScan framework with LSTM , Logistic Regression , BERT , NER	Information extraction from receipts and invoices	ICDAR 2019 SROIE	RNN, BLSTM, BERT	Receipts and invoices	-	Avg.Precision: 90.8% Soft Avg.Precision: 97.2%
[22] RNN, R-CNN	Text detection for multi-lingual text documents	Real-world Chinese passport and medical receipt	OCR, CNN	Passport and medical receipt Textual content	-	Accuracy: 94.25%
[90] RNN, LSTM, OCR	Extraction and digitization of the information from documents	3,26,471 invoice documents	CloudScan text extraction	Text from Invoices	-	F1: 0.840 Precision: 0.879 Recall: 0.804
[109] NER, OCR, BLSTM, BERT	Extracting texts from financial and biomedical documents efficiently	BC2GM BioNLP09	Word2Vec, Word Character BERT	Text from financial and biomedical documents	-	Financial Documents precision: 0.96 Biomedical

AI, organizations facilitate highly innovative levels of data processing and automation.

G. SUMMARY OF REVIEWED STUDIES

Table 14. summarizes the literature reviewed by considering different parameters such as the purpose of the study, the techniques used in the study, the feature extraction methods used in the study and few other aspects. The summary of literature reviewed is also represented in Figure 18, discussed in brief as below:

- Various feature extraction techniques are found in the literature. One-Hot-Encoding, Word2Vec, TF-IDF, GloVe, BoW, word embedding, CNN and BERT are popularly used feature extraction techniques. The features extracted using any of these techniques are passed for NER model training.
- Various ML algorithms like KNN, SVM, Naive Bays, and Neural Networks are used for text recognition and classification in OCR.
- Plenty of AI-based algorithms/techniques for automatic information extraction from unstructured documents were proposed in the literature. Few widely used approaches are LSTM and GRU, BiLSTM and CRF, CNN, RNN and BERT.

VI. DISCUSSION

In this SLR, we reviewed a large number of research papers on automatic information extraction from unstructured documents. By conducting this review, we are able to answer some key questions about various approaches used for automatic information extraction from unstructured documents found in the existing literature. We found that-

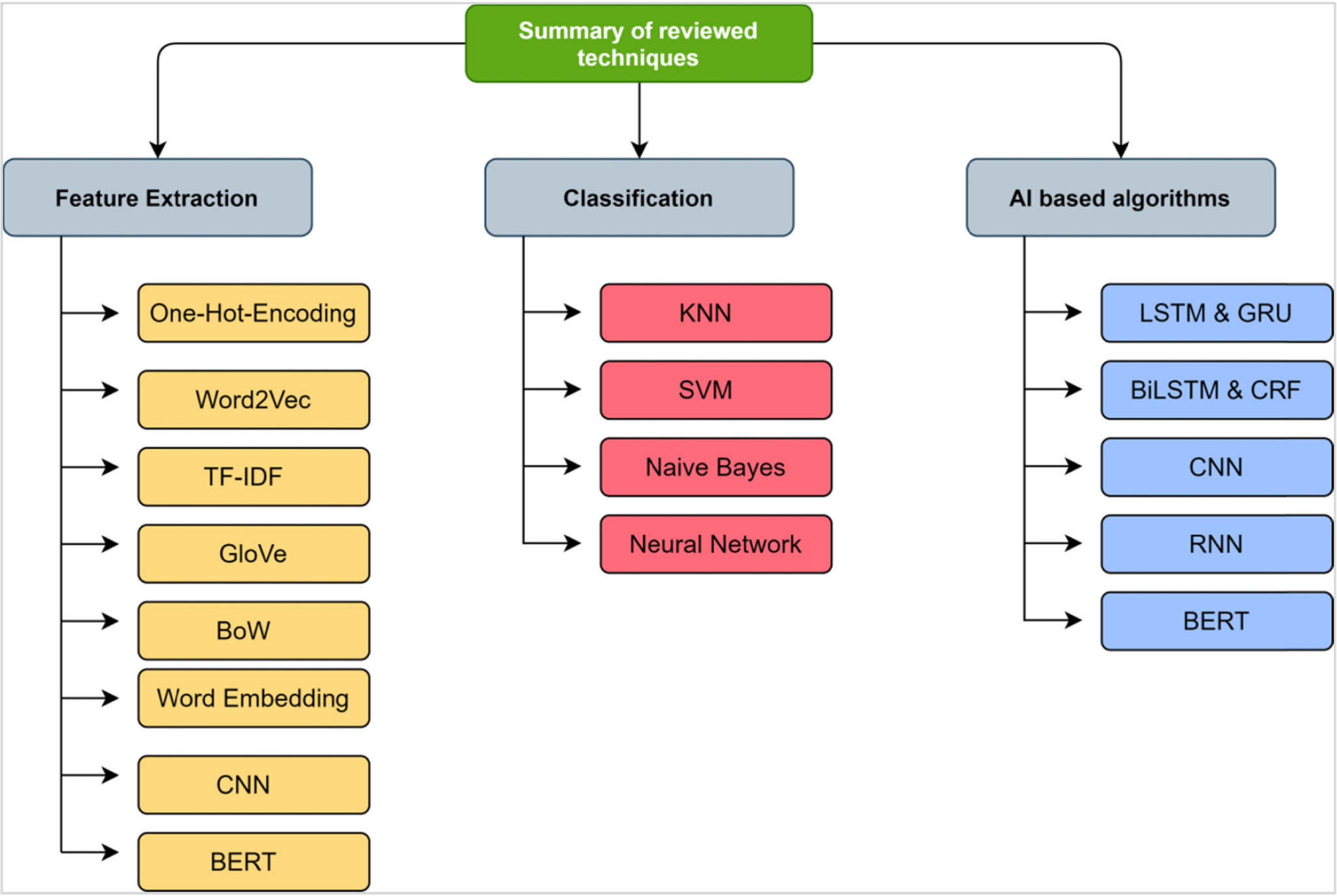


FIGURE 18. Summary of literature reviewed.

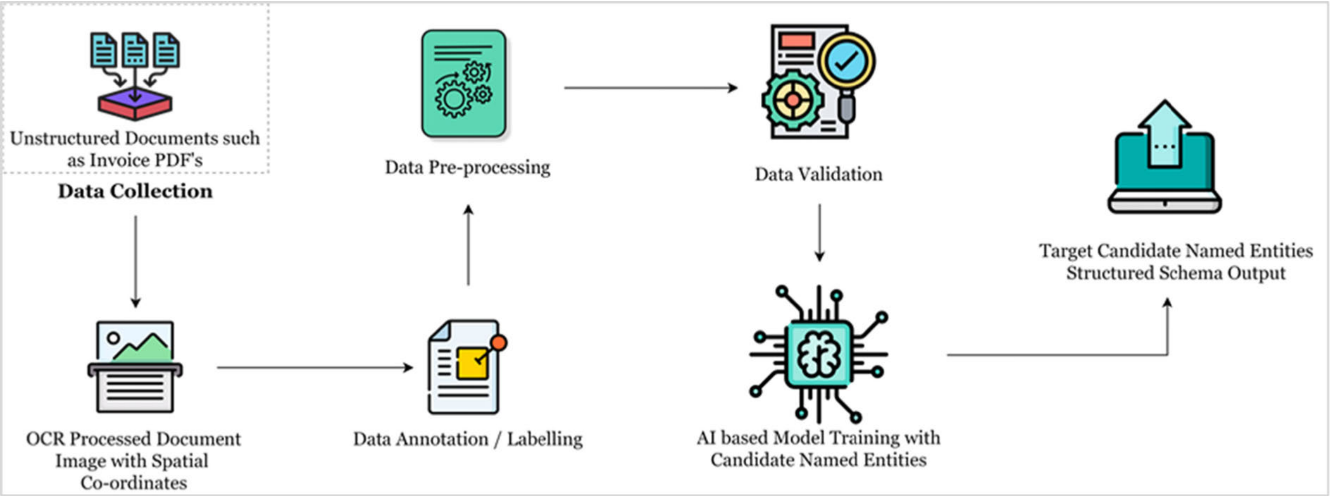

Unstructured Documents such as Invoice PDF's
Data Collection
Data Pre-processing
Data Validation
Target Candidate Named Entities Structured Schema Output
OCR Processed Document Image with Spatial Co-ordinates
Data Annotation / Labelling
AI based Model Training with Candidate Named Entities

FIGURE 19. Proposed framework architecture.

- Majority of the previous works in this domain depend on templates of the unstructured documents such as invoices, but since most of the time, different companies have different templates for their official documents like invoices. Therefore, the solutions mentioned in the previous works are not scalable over a variety of unstructured documents.
- Also, some previous works focusing on template-free solutions are solely based on the text present in the image and their sequences which is again dependent on the OCR that is used or is based on the positions of those texts in the image. These solutions have shown good results, but they can be improved if we integrate AI-based methodologies with OCR.

- To address the need of template-free AI-based model, different features of the document, such as semantic relationships, positional relationships between the named entities present in the document can be utilized.
- Data validation is an essential and crucial step in Machine Learning, but most of the literature lacks the details of these techniques.
- High-quality publicly available unstructured documents dataset for automatic information extraction task is need of time.

Subsection A. of the discussion proposes the framework for automatic information extraction from unstructured document as shown in Figure 19, which provides the solution to each of the research questions.

RQ1. What are the different challenges with information extraction techniques to deal with unstructured data?

Data produced through the daily working of an organization, such as emails, PowerPoint presentations, Word documents, PDF, images, and audio-visual records, contribute to the massive volume of the unstructured data. The researchers in multiple studies described the term unstructured data with the help of 3 V's, that is, Velocity, Volume, and Variety. The unstructured data is not well organized or easy to access. The organizations have started realizing the benefits of analyzing and integrating this data into their information management system. It may improve their productivity significantly. The analysis can also provide certain information to help the businesses to make the crucial decisions. However, to analyze and manage the unstructured data effectively, the organizations have to pay a higher cost. Traditional information extraction techniques are template-based or rule-based. Defining rules or providing a template for each new and diverse document type poses a big challenge for existing information extraction techniques. Existing information extraction techniques face certain challenges to deal with complex and multiple layout documents such as invoices, purchase orders in real-world scenario. We discussed few unstructured data issues like data sparsity, poor morphology, multiple language vocabulary, lack of quantity and quality data, non-standard phrases, domain and language-dependent entities in Section V. We observed that, AI-powered information extraction techniques have a strong potential to deal with such unstructured data challenges. Researchers in this area can explore and use AI in various domains and increase the productivity. As the unstructured data growth is exponential, this research area has the significant scope and numerous opportunities to analyze and manage the unstructured data.

RQ2. What are the different datasets available for unstructured data processing?

Most of the datasets used in the research studies are domain-dependent and language-specific. From the literature studied, it is observed that to develop a general purpose information extraction model, a comprehensive dataset containing common, general- purpose and normalized entity annotations is a prerequisite. Data sparsity such as diversity in language vocabulary, various acronyms, various rules to remove language ambiguity, and multiple languages pose challenges for the existing information extraction techniques. The existing literature lacks the most representative, comprehensive, and heterogeneous publicly available dataset, which is domain-independent. The size of the dataset is also an issue observed in the research. Lack of large labeled corpus is also a concern for this research area. The quality of scanned document images in most publicly available datasets has low-resolution, leading to poor OCR results. Datasets also consists of some missing data values, noisy, and skewed images. Use of such datasets produces less valuable and meaningless extractions. Most of the self-built/custom document datasets are confidential or sensitive as they contain private information about individuals, administrations, or companies. We discussed the issues and challenges with the existing publicly available datasets in Section V.

RQ3. What are different data validation techniques used for the quality assessment of data?

AI-based models learn to differentiate and act on complex patterns in data without explicit programming. The algorithm learns by using huge amounts of training data to fine-tune its internal parameters until it can reliably differentiate similar patterns in data it has not seen before. The AI-based model is susceptible to the quality of the data. Therefore, it is essential to evaluate the data quality by some statistical measure. Various statistical techniques are used to obtain the high-quality training data. Feature selection with Chi-square is one of the methods used for data quality improvement. It aims to select the most relevant input feature variables by calculating the dependence of the input variable on the output variable. Most of the information extraction dataset involves self-built dataset with manual annotations. Several annotators are involved in annotating the big dataset. The performance or correctness of the annotators in labeling the dataset is evaluated based on Cohen's Kappa statistical measure. The κ coefficient in Cohen's Kappa inter-rater reliability/agreement is used to measure document quality and decide the agreement rate between two annotators. The κ coefficient values in Cohen's Kappa are represented as: κ coefficient values ≤ 0 as indicating no agreement and 0.01–0.20 as none to slight, 0.21–0.40 as fair, 0.41– 0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1.00 as an almost perfect agreement. Another quality assessment test is K-Fold cross-validation. It determines which classification algorithm should be used, where the training data is divided into several parts known as folds, and several iterations are run. In each iteration, one fold is considered as testing data and the remaining as training data. This process is continued until the last fold is considered testing data, and the score of each iteration is noted down. Thus, the goal is to identify the best-suited data samples and improve data quality using K-fold cross-validation. Generally, for testing the data quality, the value $K = 5$ or 10 is usually preferred, which allows for choosing better samples. Refer Section V for detail discussion on data validation techniques.

RQ4. What are the different AI approaches used for unstructured data processing?

An efficient utilization of the unstructured data is a tedious, time-taking, laborious, and erroneous task. The information extraction techniques help in valuable and insightful information extraction from this data, making it useful for further analysis. Our SLR presents several information extraction techniques and their comparative analysis. It is observed that various information extraction methods are used in the perspective of diverse domains and tasks. Hence, the information extraction techniques are categorized in the present study to achieve an automatic information extraction of business documents efficiently. It is quite challenging to decide on a standard information extraction method for different application areas and complex or multiple layouts of the unstructured documents. It is considered the biggest challenge in the information extraction research. Traditional approaches for the unstructured data analysis have focused on rule-based (RPA), or template-based (OCR) approaches, which are time-consuming and expensive to implement. So, the researchers and businesses were started looking towards a solution that integrates AI-enabled algorithms to process the unstructured documents. AI-based solutions possess CV and NLP capabilities combined with RPA or OCR workflows, to provide an end-to-end automation solutions. The AI-based automatic information extraction in document processing has tremendous significance in the applications like banking, healthcare, financial sectors, and other domains. Extracting valuable and accurate information automatically from unstructured documents is a significant and essential task in NLP systems. Various application areas are utilizing an automated information extraction techniques for different purposes. We discussed a few of these areas along with the task in Section I. Identifying and classifying named entities from unstructured documents, that is, NER, is a specific information extraction task. There are various approaches available to tackle the NER task, which we have discussed in detail in Section V. Recently, this area is conquered by DL techniques. Automatic feature learning makes Deep Learning a powerful technique to solve automatic information extraction problem. DL models are competent to perform end-to-end tasks, from automatic feature extraction to the final classification. The only challenge nowadays in Deep Learning is choosing or creating the right Neural Network architecture suitable to perform a specific task, selecting an appropriate cost function, and gathering a lot of training data.

RQ5. What are the different online tools available for automatic information extraction from unstructured document?

Very few studies have reported the tools or frameworks developed for information extraction tasks for various domains. The tools or frameworks developed are restricted to public usage, which means they are either commercial tools or designed only for business organizations themselves. We discussed few interesting tools such as CloudScan, CUTIE, and few others in Section V. The literature lacks the details of the steps involved in developing these tools or frameworks and their dataset used, techniques, and evaluation metrics.

A. PROPOSED FRAMEWORK

We began the SLR with the objective to answer five RQs to build the foundation for our future research. These guiding questions led the foundation for our proposed framework for automatic information extraction from unstructured document processing. Based on the findings of our RQs, we now have the necessary information to propose an innovative new framework. Our proposed framework has six steps as shown in Figure 19. These are as follows:

- **Dataset Collection:** Unstructured documents such as scanned PDF of invoices from different suppliers with varied and complex layouts will be gathered for training our model on complex layouts.
- **OCR Processed Document Image:** The text present in the image is detected and extracted using an Optical Character Recognition engine (such as OCR Google Vision API). There are various OCR engines available which researchers may use for their specific task. Google Vision API supports a wide range of languages, providing the automatic identification of language. Every text annotation field has vertices (XY co-ordinates) that outline the position of the recognized element on the document. By combining spatial co-ordinates and semantics of entities, we will improve the learning capacity of a model. This also provides guidance for future work for us and other researchers by performing text extraction based on relative spatial co-ordinates making our model template independent.
- **Data Annotations/Labeling:** After getting the text from the previous step, these text files are fed into an annotation tool (such as UBIAI), where various target entities such as Buyer name, Invoice number, Invoice date, and GST number are manually annotated.
- **Data Pre-processing:** Then, various pre-processing steps will be performed on the data. These include removal of stop-words, lowercase conversion of all the text data, removal of non-alphanumeric characters, and removal of blank rows and joint words. The noise present in the dataset, such as incorrect text will be manually removed. Developing high-quality dataset with few advanced pre-processing techniques which would be used by other researchers freely is another future direction.
- **Data Validation:** In self-built datasets, it is necessary to validate the dataset before building any model on it. Data validation is important to check and improve the quality of training data. Suitable and potential statistical test will be carried out on the training data to ensure the quality of data by finding the level of significance (p-value) of features. This statistical test will be used for data validation. Exploring such statistical tests and

validating the data quality before model training, to get high performance model is one of the imperative future advancements.

- **AI-based Model training with a candidate named entities:** Now, to perform NER task on unseen documents, AI-based models such as BiLSTM and BERT are trained on the annotated dataset. These models are basically built on RNN, which is helpful to train sequential data. A hybrid model by combining different models can be implemented to increase the performance of a model. Developing an AI-based hybrid model for automatic information extraction from unstructured documents with complex layouts is a promising future direction.

VII. LIMITATIONS OF THE STUDY

Our SLR reviewed and critically analyzed the current information extraction techniques in-depth and provided comparative analysis and challenges to process the unstructured data. Several task-specific standard datasets, self-built datasets, and their validation methods are discussed in the study. However, limited literature studies and work in this area and diverse unstructured data formats have made the literature search and selection a time-consuming, laborious, and challenging job. Keywords to search the useful articles and techniques, whereby numerous research studies presenting various methods in the perspective of the type of the unstructured data, availability, and quality of datasets, different information extraction techniques, and tools used for information extraction, may vary or change for satisfying the defined inclusion and exclusion criteria.

One of the key limitations regarding the domain in which this SLR is outlined is that although we followed a systematic way of conducting a review, it is not assured that all the relevant works in this domain are retrieved. Regarding the search databases used, the most relevant electronic databases in the field of computer science were included. Other well-known search databases may be included for conducting SLR. Another limitation is the authors' bias about the whole data extraction process, although few quality assessment criteria were defined to reduce the effects of bias in the inclusion stage of the SLR.

The proposed framework is still in the initial planning phase, and its experimental findings are out of scope of this SLR, but it shows our future research plan and research direction.

VIII. FUTURE WORK AND OPPORTUNITIES

We will now discuss the prospective future directions for automated information extraction from unstructured documents. We followed our literature taxonomy theme as shown in Figure 7, to discuss the future research directions.

A. OPTICAL CHARACTER RECOGNITION (OCR)

Further possible research advancements in Optical Character Recognition (OCR) are discussed below:

- OCR is mainly a template-dependent approach used for text recognition and extraction. That means for every type of form or document, a unique template needs to be created. OCR is also a region-based approach, which means it extracts the text content from a user-defined region of interest. If the region is not defined, it digitizes the entire document content. Suppose the user wants only 10 data elements from a 50-page document and does not know the location of those data elements within the document. In that case, the user needs to digitize the entire 50 pages and look for those elements. In short, OCR cannot localize or contextualize the data user need. This leads to serious limitations in terms of human involvement and requires a lot of programming to get the data user need. These challenges possibly will provide direction for more advanced future research in OCR.
- OCR is restricted to primitive character extraction for digitization. Nowadays, organizations need end-to-end automation beyond digitization. Building an automation system on top of inconsistent OCR methods with advanced Machine Learning capabilities is challenging. Thus, a new future research direction is developing suitable AI-based technique with OCR that will lead to effective and scalable automation.
- One of the key shortcomings of OCR is the inaccuracy of processing multi-format unstructured documents. Designing hybrid model that offers high flexibility in processing multi-format documents is a promising future research area. Moreover, as a future work, the research in this domain can move beyond file type limitations –TIFF, JPEG, PDF, or any image file format.
- OCR outcomes depend on the quality of input data. Appropriate "text segmentation" methods and removal of noise from the background gives improved results. In the real world unstructured document formats, this is not true every time, so multiple pre-processing techniques need to be used for OCR to give better results. Advancements in these pre-processing techniques is another challenging future research focus.

B. ROBOTICS PROCESS AUTOMATION (RPA)

Further possible research advancements in Robotics Process Automation (RPA) are discussed below:

- RPA is mainly a rule-based approach used to design a "software bot" to perform a repetitive and high-volume task. The literature on RPA discusses the standardization of processes before RPA implementation and its role in providing RPA solutions. However, the different factors affecting their standardization to the RPA flexibility are areas for further future research.
- Furthermore, AI-based techniques such as CV and NLP has emerged in the automation domain. RPA can be combined with these AI-based techniques. This implausible evolution proposes a critical move in overall organizational strategy toward automating the specific

business processes and reducing human workforce for performing repetitive tasks that can be achieved proficiently and precisely by software bots. Future research in this direction would result in highly useful solutions to the organizations.

- Another important automation approach for future research is to critically understand and use RPA and Business Process Management to complement each other to scale automation across complete business processes.

C. NAMED ENTITY RECOGNITION (NER)

Further possible research advancements in Named Entity Recognition (NER) are discussed below:

1) DATASET PREPARATION

- One of the significant problems researchers face in the automatic information extraction tasks is getting suitable and good quality datasets. It is essential to obtain meaningful and valuable insights from the dataset, which can be further utilized for prediction and pattern-finding tasks. To deal with any document layout without providing a specific template, model training on documents having variations in the layout, is a prerequisite. Researchers can consider this a future opportunity to make the model more robust and scalable by creating a heterogeneous and comprehensive dataset. Creating high quality dataset comprising near real-world standard and layout documents (such as invoices from different suppliers having different layouts) with blend of proficient data cleaning and quality improvement techniques for automated information extraction research is promising future research direction.
- Combining spatial co-ordinates or visual features with semantics of entities with information extraction techniques is another future research for unstructured documents with complex layouts.
- The detailed discussion on the information extraction techniques using AI shows that data pre-processing is primarily critical to the efficiency of the information extraction techniques. So, exploring few data quality improvement methods which improves the performance of the information extraction techniques is another future research focus.

2) DATA VALIDATION

- Quality assessment of data plays an essential role in model accuracy and performance. Existing literature lacks the details on data validation methods. So, the researchers may explore and can write reviews specifically on various data validation methods for the task-specific applications. It would be useful for other researchers to understand the importance of data validation and know the availability of different data validation methods in detail.
- The fascinating future exploration is investigating and analyzing the results obtained from the different statistical data validation tests for nominal/categorical data.

3) DEVELOPING AI-BASED FRAMEWORK

- The deployment of AI with OCR and RPA is a promising future direction that can provide scalability and flexibility in automatic unstructured document processing tasks.
- Developing Hybrid models which can contribute well in fulfilling the template-free end-to-end automation solution requirements is another progressive future research focus.

IX. CONCLUSION

The SLR aims to explore the recent information extraction techniques for unstructured document processing to identify opportunities for advancements in this area. Guidelines proposed by Kitchenham and Charters were adhered to conduct the literature search for this SLR. Based on inclusion, exclusion criteria, and quality assessment criteria, 83 potential studies were finally selected to answer the research questions. It can be concluded from Figure 5. that there is a substantial rise in the publication contributions by the researchers in the last ten years. It demonstrates the importance and advancements in this research area.

The SLR extensively reviews and evaluates automatic information extraction research by-

- Identifying the challenges with the existing information extraction techniques to deal with unstructured documents processing.
- Identifying the need of developing a high-quality unstructured document dataset that is publicly available.
- Identifying available data validation methods for data quality assessment.
- Exploring this area of research for various application domains such as biomedical entity extraction, clinical named entity extraction, legal sector clause extraction, invoices extraction and few other.
- Reviewing the methods for text detection, preprocessing and recognition.
- Underlining the challenges and research opportunities in an automatic information extraction for the unstructured documents using various AI-based techniques.

The findings show that combining different techniques such as DL and NLP, called the Hybrid model, receives special attention from the researchers due to its efficiency in handling extensive unstructured data. We also observed that less attention is given to the template-free approaches to process the complex and multiple layouts of unstructured documents. Thus, developing a template-free AI-based model for automatic extraction of useful information from unstructured documents with complex and varied layout is a promising future opportunity. Our review highlights the opportunities for research in the area of OCR, RPA, and AI-based techniques used for automatic information extraction from unstructured documents.

The proposed framework aims to build the high-quality unstructured document datasets with varied and complex layouts from multiple sources, such as invoices from different suppliers, that will be publicly available to enhance future research in this domain. It helps researchers to validate the quality of data before model training with different statistical techniques, resulting in better model performance. The proposed framework further aims to develop an AI-based template-free framework for automatic information extraction from unstructured documents.

This study has several practical/industry implications for automatic information extraction adoption in the finance and legal sectors. Our results indicate that although automatic information extraction adoption has started in several other industries, additional improvements are necessary to achieve automatic information extraction from complex and varied unstructured documents. The benefits of automatic information extraction adoption are fairly clear; however, organizations have some significant challenges to address in the future with diverse and complex unstructured documents. Large organizations can leverage their position to create a first-mover advantage in the end-to-end automation for information extraction from unstructured documents, which will further strengthen their position in the automation implementation.

REFERENCES

[1] K. Adnan and R. Akbar, "An analytical study of information extraction from unstructured and multidimensional big data," *J. Big Data*, vol. 6, no. 1, p. 91, 2019.

[2] S. Burnett, D. Analyst, A. Verma, S. Analyst, and P. Srinivasan, "Unstructured data process automation a deep dive into the role of artificial intelligence (AI) in automating content-centric processes," Dallas, TX, USA, 2019.

[3] *30 Eye-Opening Big Data Statistics for 2020: Patterns are Everywhere*. Accessed: Dec. 5, 2020. [Online]. Available: https://kommandotech.com/statistics/big-data-statistics/

[4] *Big Data—Statistics & Facts | Statista*. Accessed: Dec. 5, 2020. [Online]. Available: https://www.statista.com/topics/1464/big-data/

[5] A. Masood and A. Hashmi, *Cognitive Computing Recipes*. New York, NY, USA: Apress, 2019, doi: 10.1007/978-1-4842-4106-6.

[6] K. Adnan and R. Akbar, "Limitations of information extraction methods and techniques for heterogeneous unstructured big data," *Int. J. Eng. Bus. Manag.*, vol. 11, pp. 1–23, 2019, doi: 10.1177/1847979019890771.

[7] R. K. Subudhi and B. Sahu, "A novel noise reduction method for OCR system 1," *Int. J. Comput. Sci. Technol.*, vol. 8491, pp. 82–86, 2014.

[8] J. Siderska, "Robotic process automation—A driver of digital transformation?" *Eng. Manage. Prod. Services*, vol. 12, no. 2, pp. 21–31, Jun. 2020, doi: 10.2478/emj-2020-0009.

[9] *Evolution of Robotic Process Automation (RPA): The Path to Cognitive RPA | by AIMDek Technologies | Medium*. Accessed: Dec. 6, 2020. [Online]. Available: https://medium.com/@AIMDekTech/evolution-of-robotic-process-automation-the-path-to-cognitive-rpa-c3bd52c8b865

[10] P. Martins, F. Sa, F. Morgado, and C. Cunha, "Using machine learning for cognitive robotic process automation (RPA)," in *Proc. 15th Iberian Conf. Inf. Syst. Technol. (CISTI)*, Jun. 2020, pp. 1–6, doi: 10.23919/CISTI49556.2020.9140440.

[11] M. P. Bach, Ž. Krstic, S. Seljan, and L. Turulja, "Text mining for big data analysis in financial sector: A literature review," *Sustainability*, vol. 11, no. 5, p. 1277, Feb. 2019, doi: 10.3390/su11051277.

[12] T. Al-Moslmi, M. G. Ocana, A. L. Opdahl, and C. Veres, "Named entity extraction for knowledge graphs: A literature overview," *IEEE Access*, vol. 8, pp. 32862–32881, 2020, doi: 10.1109/ACCESS.2020.2973928.

[13] Y. Wang, L. Wang, M. Rastegar-Mojarad, S. Moon, F. Shen, N. Afzal, S. Liu, Y. Zeng, S. Mehrabi, S. Sohn, and H. Liu, "Clinical information extraction applications: A literature review," *J. Biomed. Informat.*, vol. 77, pp. 34–49, Jan. 2018, doi: 10.1016/j.jbi.2017.11.011.

[14] R. Syed, S. Suriadi, M. Adams, W. Bandara, S. J. J. Leemans, C. Ouyang, A. H. M. ter Hofstede, I. van de Weerd, M. T. Wynn, and H. A. Reijers, "Robotic process automation: Contemporary themes and challenges," *Comput. Ind.*, vol. 115, Feb. 2020, Art. no. 103162, doi: 10.1016/j.compind.2019.103162.

[15] B. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering," Dept. Eng., Keele Univ., Durham Univ., Keele, U.K., Tech. Rep. EBSE-2007-01, 2007. [Online]. Available: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.117.471&rep=rep1&type=pdf

[16] J. Memon, M. Sami, R. A. Khan, and M. Uddin, "Handwritten optical character recognition (OCR): A comprehensive systematic literature review (SLR)," *IEEE Access*, vol. 8, pp. 142642–142668, 2020, doi: 10.1109/ACCESS.2020.3012542.

[17] I. Supriana and A. Nasution, "Arabic character recognition system development," *Procedia Technol.*, vol. 11, pp. 334–341, Jan. 2013, doi: 10.1016/j.protcy.2013.12.199.

[18] S. Prum, "Text-zone detection and rectification in document images captured by smartphone," in *Proc. 1st EAI Int. Conf. Comput. Sci. Eng.*, 2017, pp. 1–10.

[19] A. Kaur, S. Baghla, and S. Kumar, "Study of various character segmentation techniques for handwritten off-line cursive words: A review," *Int. J. Adv. Sci. Eng. Technol.*, vol. 3, no. 3, pp. 154–158, 2015. [Online]. Available: http://www.iraj.in/journal/journal_file/journal_pdf/6-162-1440573382154-158.pdf

[20] P. Sahare and S. B. Dhok, "Multilingual character segmentation and recognition schemes for Indian document images," *IEEE Access*, vol. 6, pp. 10603–10617, 2018, doi: 10.1109/ACCESS.2018.2795104.

[21] W. Liu, Y. Zhang, and B. Wan, "Unstructured document recognition on business invoice," Mach. Learn., Stanford iTunes Univ., Stanford, CA, USA, Tech. Rep., 2016. [Online]. Available: http://cs229.stanford.edu/proj2016/report/LiuWanZhang-UnstructuredDocumentRecognitionOnBusinessInvoice-report.pdf

[22] Y. Ye, S. Zhu, J. Wang, Q. Du, Y. Yang, D. Tu, L. Wang, and J. Luo, "A unified scheme of text localization and structured data extraction for joint OCR and data mining," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2018, pp. 2373–2382, doi: 10.1109/BigData.2018.8622129.

[23] M. Kanya and T. Ravi, "Named entity recognition from biomedical text–an information extraction task," *ICTACT J. Soft Comput.*, vol. 6, no. 4, pp. 1303–1307, Jul. 2016, doi: 10.21917/ijsc.2016.0179.

[24] C. Reul, D. Christ, A. Hartelt, N. Balbach, M. Wehner, U. Springmann, C. Wick, C. Grundig, A. Büttner, and F. Puppe, "OCR4all—An open-source tool providing a (semi-)automatic OCR workflow for historical printings," *Appl. Sci.*, vol. 9, no. 22, p. 4853, Nov. 2019, doi: 10.3390/app9224853.

[25] F. Santos, R. Pereira, and J. B. Vasconcelos, "Toward robotic process automation implementation: An end-to-end perspective," *Bus. Process Manage. J.*, vol. 26, no. 2, pp. 405–420, Sep. 2019, doi: 10.1108/BPMJ-12-2018-0380.

[26] M. Kukreja, "Study of robotic process automation (RPA)," *Int. J. Recent Innov. Trends Comput. Commun.*, vol. 4, no. 6, pp. 434–437, Jun. 2016.

[27] A. Wróblewska, T. Stanisławek, B. Prus-Zajaczkowski, and Ł. Garncarek, "Robotic process automation of unstructured data with machine learning," in *Proc. Position Papers Federated Conf. Comput. Sci. Inf. Syst.*, vol. 16, Sep. 2018, pp. 9–16, doi: 10.15439/2018f373.

[28] J. M. Steinkamp, W. Bala, A. Sharma, and J. J. Kantrowitz, "Task definition, annotated dataset, and supervised natural language processing models for symptom extraction from unstructured clinical notes," *J. Biomed. Informat.*, vol. 102, Feb. 2020, Art. no. 103354, doi: 10.1016/j.jbi.2019.103354.

[29] N. Perera, M. Dehmer, and F. Emmert-Streib, "Named entity recognition and relation detection for biomedical information extraction," *Frontiers Cell Develop. Biol.*, vol. 8, p. 673, Aug. 2020, doi: 10.3389/fcell.2020.00673.

[30] F. Brauer, R. Rieger, A. Mocan, and W. M. Barczynski, "Enabling information extraction by inference of regular expressions from sample entities," in *Proc. 20th ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*, 2011, pp. 1285–1294, doi: 10.1145/2063576.2063763.

[31] B. Moysset, C. Kermorvant, and C. Wolf, "Learning to detect, localize and recognize many text objects in document images from few examples," *Int. J. Document Anal. Recognit.*, vol. 21, no. 3, pp. 161–175, Sep. 2018, doi: 10.1007/s10032-018-0305-2.

[32] A. D. Le, D. V. Pham, and T. A. Nguyen, "Deep learning approach for receipt recognition," in *Future Data and Security Engineering* (Lecture Notes in Computer Science: Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 11814. Springer, 2019, pp. 705–712.

[33] X. Zhao, E. Niu, Z. Wu, and X. Wang, "CUTIE: Learning to understand documents with convolutional universal text information extractor," 2019, *arXiv:1903.12363*. [Online]. Available: http://arxiv.org/abs/1903.12363

[34] J. Yang, Y. Liu, M. Qian, C. Guan, and X. Yuan, "Information extraction from electronic medical records using multitask recurrent neural network with contextual word embedding," *Appl. Sci.*, vol. 9, no. 18, p. 3658, Sep. 2019, doi: 10.3390/app9183658.

[35] C. Artaud, A. Doucet, J.-M. Ogier, V. P. D'andecy, and V. Poulain. *Receipt Dataset for Fraud Detection*. Accessed: Sep. 21, 2020. [Online]. Available: https://hal.archives-ouvertes.fr/hal-02316349.

[36] A. S. Tarawneh, A. B. Hassanat, D. Chetverikov, I. Lendak, and C. Verma, "Invoice classification using deep features and machine learning techniques," in *Proc. IEEE Jordan Int. Joint Conf. Electr. Eng. Inf. Technol. (JEEIT)*, Apr. 2019, pp. 855–859, doi: 10.1109/JEEIT.2019.8717504.

[37] C. Pitou and J. Diatta, "Textual information extraction in document images guided by a concept lattice," in *Proc. CEUR Workshop*, vol. 1624, 2016, pp. 325–336.

[38] A. Singh and S. Desai, "Optical character recognition using template matching and back propagation algorithm," in *Proc. Int. Conf. Inventive Comput. Technol. (ICICT)*, Aug. 2016, pp. 1–6, doi: 10.1109/INVENTIVE.2016.7830161.

[39] H. Sidhwa, S. Kulshrestha, S. Malhotra, and S. Virmani, "Text extraction from bills and invoices," in *Proc. Int. Conf. Adv. Comput., Commun. Control Netw. (ICACCCN)*, Oct. 2018, pp. 564–568, doi: 10.1109/ICACCCN.2018.8748309.

[40] *Five Case Studies to Inspire Your Intelligent Automation Strategy*, Kofax, Irvine, CA, USA, 2019.

[41] D. Šimek and R. Šperka, "How robot/human orchestration can help in an HR department: A case study from a pilot implementation," *Organizacija*, vol. 52, no. 3, pp. 204–217, Aug. 2019, doi: 10.2478/orga-2019-0013.

[42] P. Shah, S. Joshi, and A. K. Pandey, "Legal clause extraction from contract using machine learning with heuristics improvement," in *Proc. 4th Int. Conf. Comput. Commun. Autom. (ICCCA)*, Dec. 2018, pp. 1–3, doi: 10.1109/CCAA.2018.8777602.

[43] D. Chakrabarti, N. Patodia, U. Bhattacharya, I. Mitra, S. Roy, J. Mandi, N. Roy, and P. Nandy, "Use of artificial intelligence to analyse risk in legal documents for a better decision support," in *Proc. TENCON-IEEE Region 10th Conf.*, Oct. 2018, pp. 683–688, doi: 10.1109/TENCON.2018.8650382.

[44] S. Joshi, P. Shah, and A. K. Pandey, "Location identification, extraction and disambiguation using machine learning in legal contracts," in *Proc. 4th Int. Conf. Comput. Commun. Autom. (ICCCA)*, Dec. 2018, pp. 1–5, doi: 10.1109/CCAA.2018.8777631.

[45] I. Chalkidis, I. Androutsopoulos, and A. Michos, "Extracting contract elements," in *Proc. 16th Ed. Int. Conf. Articial Intell. Law*, Jun. 2017, pp. 19–28, doi: 10.1145/3086512.3086515.

[46] I. Chalkidis and I. Androutsopoulos, "A deep learning approach to contract element extraction," in *Frontiers in Artificial Intelligence and Applications*, vol. 302. IOS Press, 2017, pp. 155–164.

[47] Y. Sun, X. Mao, S. Hong, W. Xu, and G. Gui, "Template matching-based method for intelligent invoice information identification," *IEEE Access*, vol. 7, pp. 28392–28401, 2019, doi: 10.1109/ACCESS.2019.2901943.

[48] S. Patel and D. Bhatt, "Abstractive information extraction from scanned invoices (AIESI) using end-to-end sequential approach," 2020, *arXiv:2009.05728*. [Online]. Available: http://arxiv.org/abs/2009.05728

[49] Y. Chen, J. E. Argentinis, and G. Weber, "IBM Watson: How cognitive computing can be applied to big data challenges in life sciences research," *Clin. Therapeutics*, vol. 38, no. 4, pp. 688–701, Apr. 2016, doi: 10.1016/j.clinthera.2015.12.001.

[50] S. Purushotham, C. Meng, Z. Che, and Y. Liu, "Benchmarking deep learning models on large healthcare datasets," *J. Biomed. Informat.*, vol. 83, pp. 112–134, Jul. 2018, doi: 10.1016/j.jbi.2018.04.007.

[51] D. Tkaczyk, P. Szostek, M. Fedoryszak, P. J. Dendek, and Ł. Bolikowski, "CERMINE: Automatic extraction of structured metadata from scientific literature," *Int. J. Document Anal. Recognit.*, vol. 18, no. 4, pp. 317–335, Dec. 2015, doi: 10.1007/s10032-015-0249-8.

[52] D. D. A. Bui, G. D. Fiol, and S. Jonnalagadda, "PDF text classification to leverage information extraction from publication reports," *J. Biomed. Informat.*, vol. 61, pp. 141–148, Jun. 2016, doi: 10.1016/j.jbi.2016.03.026.

[53] A. C. Eberendu, "Unstructured data: An overview of the data of big data," *Int. J. Comput. Trends Technol.*, vol. 38, no. 1, pp. 46–50, Aug. 2016, doi: 10.14445/22312803/ijctt-v38p109.

[54] G. Zaman, H. Mahdin, and K. Hussain, "Information extraction from semi and unstructured data sources: A systematic literature review," *ICIC Exp. Lett.*, vol. 14, no. 6, pp. 593–603, Jun. 2020, doi: 10.24507/icicel.14.06.593.

[55] R. Alfred, L. C. Leong, C. K. On, and P. Anthony, "Malay named entity recognition based on rule-based approach," *Int. J. Mach. Learn. Comput.*, vol. 4, no. 3, pp. 300–306, Jun. 2014, doi: 10.7763/ijmlc.2014.v4.428.

[56] B. Davis, B. Morse, S. Cohen, B. Price, and C. Tensmeyer, "Deep visual template-free form parsing," in *Proc. Int. Conf. Document Anal. Recognit. (ICDAR)*, Sep. 2019, pp. 134–141, doi: 10.1109/ICDAR.2019.00030.

[57] Y. S. Chernyshova, A. V. Sheshkus, and V. V. Arlazarov, "Two-step CNN framework for text line recognition in camera-captured images," *IEEE Access*, vol. 8, pp. 32587–32600, 2020, doi: 10.1109/ACCESS.2020.2974051.

[58] B. Bataineh, "A printed PAW image database of Arabic language for document analysis and recognition," *J. ICT Res. Appl.*, vol. 11, no. 2, pp. 199–211, 2017, doi: 10.5614/itbj.ict.res.appl.2017.11.2.6.

[59] C. Clausner, A. Antonacopoulos, and S. Pletschacher, "Efficient and effective OCR engine training," *Int. J. Document Anal. Recognit.*, vol. 23, no. 1, pp. 73–88, Mar. 2020, doi: 10.1007/s10032-019-00347-8.

[60] L. Todoran, M. Worring, and A. W. M. Smeulders, "The UvA color document dataset," *Int. J. Document Anal. Recognit.*, vol. 7, no. 4, pp. 228–240, Sep. 2005, doi: 10.1007/s10032-004-0135-2.

[61] A. W. Harley, A. Ufkes, and K. G. Derpanis, "Evaluation of deep convolutional nets for document image classification and retrieval," in *Proc. 13th Int. Conf. Document Anal. Recognit. (ICDAR)*, Aug. 2015, pp. 991–995, doi: 10.1109/ICDAR.2015.7333910.

[62] S. Gehrmann, F. Dernoncourt, Y. Li, E. T. Carlson, J. T. Wu, J. Welt, J. Foote, E. T. Moseley, D. W. Grant, P. D. Tyler, and L. A. Celi, "Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives," *PLoS ONE*, vol. 13, no. 2, Feb. 2018, Art. no. e0192360, doi: 10.1371/journal.pone.0192360.

[63] A. Abbas, M. Afzal, J. Hussain, and S. Lee, "Meaningful information extraction from unstructured clinical documents," in *Proc. Asia–Pacific Adv. Netw.*, vol. 48, 2019, pp. 42–47. Accessed: Sep. 17, 2020. [Online]. Available: https://www.researchgate.net/publication/336797539_Meaningful_Information_Extraction_from_Unstructured_Clinical_Documents

[64] D. Tkaczyk, P. Szostek, and L. Bolikowski, "GROTOAP2—The methodology of creating a large ground truth dataset of scientific articles," *D-Lib Mag.*, vol. 20, 2014, doi: 10.1045/november14-tkaczyk.

[65] C.-A. Boiangiu, O.-A. Dinu, C. Popescu, N. Constantin, and C. Petrescu, "Voting-based document image skew detection," *Appl. Sci.*, vol. 10, no. 7, p. 2236, Mar. 2020, doi: 10.3390/app10072236.

[66] E. L. Park, S. Cho, and P. Kang, "Supervised paragraph vector: Distributed representations of words, documents and class labels," *IEEE Access*, vol. 7, pp. 29051–29064, 2019, doi: 10.1109/ACCESS.2019.2901933.

[67] D. Christou, "Feature extraction using latent Dirichlet allocation and neural networks: A case study on movie synopses," 2016, *arXiv:1604.01272*. [Online]. Available: http://arxiv.org/abs/1604.01272

[68] B. Jang, M. Kim, G. Harerimana, S.-U. Kang, and J. W. Kim, "Bi-LSTM model to increase accuracy in text classification: Combining Word2vec CNN and attention mechanism," *Appl. Sci.*, vol. 10, no. 17, p. 5841, Aug. 2020, doi: 10.3390/app10175841.

[69] J. He, L. Wang, L. Liu, J. Feng, and H. Wu, "Long document classification from local word glimpses via recurrent attention learning," *IEEE Access*, vol. 7, pp. 40707–40718, 2019, doi: 10.1109/ACCESS.2019.2907992.

[70] L. Arras, F. Horn, G. Montavon, K.-R. Müller, and W. Samek, "What is relevant in a text document?" *PLoS ONE*, vol. 12, no. 8, pp. 1–19, 2016. [Online]. Available: http://arxiv.org/abs/1612.07843.

[71] J. C. Campbell, A. Hindle, and E. Stroulia, ''Latent Dirichlet allocation: Extracting topics from software engineering data,'' in *The Art and Science of Analyzing Software Data*. Amsterdam, The Netherlands: Elsevier, 2015, pp. 139–159.

[72] S. Eken, H. Menhour, and K. Koksal, ''DoCA: A content-based automatic classification system over digital documents,'' *IEEE Access*, vol. 7, pp. 97996–98004, 2019, doi: 10.1109/ACCESS.2019.2930339.

[73] M. Binkhonain and L. Zhao, ''A review of machine learning algorithms for identification and classification of non-functional requirements,'' *Expert Syst. Appl.*, vol. 1, Apr. 2019, Art. no. 100001.

[74] J. Huang, J. Chai, and S. Cho, ''Deep learning in finance and banking: A literature review and classification,'' *Frontiers Bus. Res. China*, vol. 14, no. 1, p. 13, Dec. 2020, doi: 10.1186/s11782-020-00082-6.

[75] Z. Huang, K. Chen, J. He, X. Bai, D. Karatzas, S. Lu, and C. V. Jawahar, ''ICDAR2019 competition on scanned receipt OCR and information extraction,'' in *Proc. Int. Conf. Document Anal. Recognit. (ICDAR)*, Sep. 2019, pp. 1516–1520, doi: 10.1109/ICDAR.2019.00244.

[76] K. Jung, K. I. Kim, and A. K. Jain, ''Text information extraction in images and video: A survey,'' *Pattern Recognit.*, vol. 37, no. 5, pp. 977–997, May 2004, doi: 10.1016/j.patcog.2003.10.012.

[77] J. I. Toledo, M. Carbonell, A. Fornés, and J. Lladós, ''Information extraction from historical handwritten document images with a context-aware neural model,'' *Pattern Recognit.*, vol. 86, pp. 27–36, Feb. 2019, doi: 10.1016/j.patcog.2018.08.020.

[78] H. T. Ha, ''Recognition of invoices from scanned documents,'' in *Proc. Recent Adv. Slavon. Nat. Lang. Process.*, 2017, pp. 71–78.

[79] T. Grüning, G. Leifert, T. Strauß, J. Michael, and R. Labahn, ''A two-stage method for text line detection in historical documents,'' *Int. J. Document Anal. Recognit.*, vol. 22, no. 3, pp. 285–302, Sep. 2019, doi: 10.1007/s10032-019-00332-1.

[80] U. Munir and M. Ozturk, ''Automatic character extraction from handwritten scanned documents to build large scale database,'' in *Proc. Sci. Meeting Elect.-Electron. Biomed. Eng. Comput. Sci. (EBBT)*, Apr. 2019, pp. 1–4, doi: 10.1109/EBBT.2019.8741984.

[81] H. Chao and J. Fan, ''Layout and content extraction for PDF documents,'' in *Document Analysis Systems VI* (Lecture Notes in Computer Science: Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 3163. Springer-Verlag, 2004, pp. 213–224.

[82] N. Nobile and C. Y. Suen, ''Text segmentation for document recognition,'' in *Handbook of Document Image Processing and Recognition*. London, U.K.: Springer, 2014, pp. 257–290.

[83] W. Xue, Q. Li, and Q. Xue, ''Text detection and recognition for images of medical laboratory reports with a deep learning approach,'' *IEEE Access*, vol. 8, pp. 407–416, 2020, doi: 10.1109/ACCESS.2019.2961964.

[84] Q. Ye and D. Doermann, ''Text detection and recognition in imagery: A survey,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 7, pp. 1480–1500, Jul. 2015, doi: 10.1109/TPAMI.2014.2366765.

[85] D. Karatzas, S. R. Mestre, J. Mas, F. Nourbakhsh, and P. P. Roy, ''ICDAR 2011 robust reading competition–challenge 1: Reading text in born-digital images (Web and Email),'' in *Proc. Int. Conf. Document Anal. Recognit.*, Sep. 2011, pp. 1485–1490, doi: 10.1109/ICDAR.2011.295.

[86] J. Johnson, A. Gupta, and L. Fei-Fei, ''Image generation from scene graphs,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1219–1228, doi: 10.1109/CVPR.2018.00133.

[87] J. Laubrock and A. Dunst, ''Computational approaches to comics analysis,'' *Topics Cognit. Sci.*, vol. 12, no. 1, pp. 274–310, Jan. 2020, doi: 10.1111/tops.12476.

[88] P. Singh, S. Varadarajan, A. N. Singh, and M. M. Srivastava, ''Multi-domain document layout understanding using few-shot object detection,'' in *Image Analysis and Recognition* (Lecture Notes in Computer Science: Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 12132. Springer, 2020, pp. 89–99.

[89] G. Mehul, P. Ankita, D. Namrata, G. Rahul, and S. Sheth, ''Text-based image segmentation methodology,'' *Procedia Technol.*, vol. 14, pp. 465–472, 2014.

[90] R. B. Palm, O. Winther, and F. Laws, ''CloudScan—A configuration-free invoice analysis system using recurrent neural networks,'' in *Proc. 14th IAPR Int. Conf. Document Anal. Recognit. (ICDAR)*, Nov. 2017, pp. 406–413, doi: 10.1109/ICDAR.2017.74.

[91] C. Clausner, S. Pletschacher, and A. Antonacopoulos, ''Flexible character accuracy measure for reading-order-independent evaluation,'' *Pattern Recognit. Lett.*, vol. 131, pp. 390–397, Mar. 2020, doi: 10.1016/j.patrec.2020.02.003.

[92] J. G. Enriquez, A. Jimenez-Ramirez, F. J. Dominguez-Mayo, and J. A. Garcia-Garcia, ''Robotic process automation: A scientific and industrial systematic mapping study,'' *IEEE Access*, vol. 8, pp. 39113–39129, 2020, doi: 10.1109/ACCESS.2020.2974934.

[93] P. Hofmann, C. Samp, and N. Urbach, ''Robotic process automation,'' *Electron. Markets*, vol. 30, no. 1, pp. 99–106, Mar. 2020, doi: 10.1007/s12525-019-00365-8.

[94] J. B. Kim, ''Implementation strategy and model of robotic process automation for green it development: An exploratory study,'' *J. Green Eng.*, vol. 10, no. 7, pp. 3559–3574, Jul. 2020.

[95] J. Wanner, A. Hofmann, M. Fischer, F. Imgrund, C. Janiesch, and J. Geyer-Klingeberg, ''Process selection in RPA projects—Towards a quantifiable method of decision making,'' in *Proc. 40th Int. Conf. Inf. Syst. (ICIS)*. Munich, Germany: Association for Information Systems, 2020.

[96] A. R. Katti, C. Reisswig, C. Guder, S. Brarda, S. Bickel, J. Höhne, and J. B. Faddoul, ''Chargrid: Towards understanding 2D documents,'' in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 4459–4469, doi: 10.18653/v1/d18-1476.

[97] G. Zhu and C. A. Iglesias, ''Exploiting semantic similarity for named entity disambiguation in knowledge graphs,'' *Expert Syst. Appl.*, vol. 101, pp. 8–24, Jul. 2018, doi: 10.1016/j.eswa.2018.02.011.

[98] K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, ''Text classification algorithms: A survey,'' *Information*, vol. 10, no. 4, pp. 1–68, 2019, doi: 10.3390/info10040150.

[99] R. Sharma, N. Goel, N. Aggarwal, P. Kaur, and C. Prakash, ''Next word prediction in Hindi using deep learning techniques,'' in *Proc. Int. Conf. Data Sci. Eng. (ICDSE)*, Sep. 2019, pp. 55–60, doi: 10.1109/icdse47409.2019.8971796.

[100] G. Rabby, S. Azad, M. Mufti, K. Z. Zamli, and M. M. Rahman, ''A flexible keyphrase extraction technique for academic literature,'' *Procedia Comput. Sci.*, vol. 135, pp. 553–563, 2018, doi: 10.1016/j.procs.2018.08.208.

[101] S. Jaf and C. Calder, ''Deep learning for natural language parsing,'' *IEEE Access*, vol. 7, pp. 131363–131373, 2019, doi: 10.1109/access.2019.2939687.

[102] S. Pyo, E. Kim, and M. Kim, ''LDA-based unified topic modeling for similar TV user grouping and TV program recommendation,'' *IEEE Trans. Cybern.*, vol. 45, no. 8, pp. 1476–1490, Aug. 2015, doi: 10.1109/TCYB.2014.2353577.

[103] M. Rhanoui, M. Mikram, S. Yousfi, and S. Barzali, ''A CNN-BiLSTM model for document-level sentiment analysis,'' *Mach. Learn. Knowl. Extraction*, vol. 1, no. 3, pp. 832–847, Jul. 2019, doi: 10.3390/make1030048.

[104] M. A. K. Oziuddeen, S. Poruran, and M. Y. Caffiyar, ''A novel deep convolutional neural network architecture based on transfer learning for handwritten Urdu character recognition,'' *Tehnicki Vjesnik*, vol. 27, no. 4, pp. 1160–1165, Aug. 2020, doi: 10.17559/TV-20190319095323.

[105] Y. Chen, J. Wang, P. Li, and P. Guo, ''Single document keyword extraction via quantifying higher-order structural features of word co-occurrence graph,'' *Comput. Speech Lang.*, vol. 57, pp. 98–107, Sep. 2019, doi: 10.1016/j.csl.2019.01.007.

[106] A. Mandelbaum and A. Shalev, ''Word embeddings and their use in sentence classification tasks,'' 2016, *arXiv:1610.08229*. [Online]. Available: http://arxiv.org/abs/1610.08229

[107] N. D. Grujic and V. M. Milovanovic, ''Natural language processing for associative word predictions,'' in *Proc. IEEE EUROCON-18th Int. Conf. Smart Technol.*, Jul. 2019, pp. 1–6, doi: 10.1109/EUROCON.2019.8861547.

[108] X. Wang, Z. Cui, L. Jiang, W. Lu, and J. Li, ''WordleNet: A visualization approach for relationship exploration in document collection,'' *Tsinghua Sci. Technol.*, vol. 25, no. 3, pp. 384–400, Jun. 2020, doi: 10.26599/TST.2019.9010005.

[109] S. Francis, J. V. Landeghem, and M.-F. Moens, ''Transfer learning for named entity recognition in financial and biomedical documents,'' *Information*, vol. 10, no. 8, p. 248, Jul. 2019, doi: 10.3390/info10080248.

[110] Y. Zhang and W. Xiao, ''Keyphrase generation based on deep Seq2seq model,'' *IEEE Access*, vol. 6, pp. 46047–46057, Aug. 2018, doi: 10.1109/ACCESS.2018.2865589.

[111] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, ''Language models are unsupervised multitask learners,'' Tech. Rep., 2018. [Online]. Available: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

[112] R. B. Palm, F. Laws, and O. Winther, "Attend, copy, parse end-to-end information extraction from documents," in *Proc. Int. Conf. Document Anal. Recognit. (ICDAR)*, Sep. 2019, pp. 329–336, doi: 10.1109/ICDAR.2019.00060.

[113] X. Ma and E. Hovy, "End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, vol. 2, 2016, pp. 1064–1074, doi: 10.18653/v1/p16-1101.

[114] F. Yi, Y.-F. Zhao, G.-Q. Sheng, K. Xie, C. Wen, X.-G. Tang, and X. Qi, "Dual model medical invoices recognition," *Sensors*, vol. 19, no. 20, p. 4370, Oct. 2019, doi: 10.3390/s19204370.

[115] S. Su, Shirabad, Matwin, and Huang. (Nov. 30, 2012). *Discriminative Multinominal Naive Bayes for Text Classification*. Accessed: Oct. 30, 2020. [Online]. Available: http://www.site.uottawa.ca/~stan/csi5387/DMNB-paper.pdf on

[116] X. Han and L. Wang, "A novel document-level relation extraction method based on BERT and entity information," *IEEE Access*, vol. 8, pp. 96912–96919, 2020, doi: 10.1109/ACCESS.2020.2996642.

[117] Y. Li, K. He, J. Sun, and others, "R-FCN: Object detection via region-based fully convolutional networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2016, pp. 379–387. [Online]. Available: http://papers.nips.cc/paper/6465-r-fcn-object-detection-via-region-based-fully-convolutional-networks.pdf

DIPALI BAVISKAR received the master's degree in computer science and engineering from the MGM College of Engineering, Nanded. She is currently pursuing the Ph.D. degree with the Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune. She is currently working as an Assistant Professor with the School of Computer Engineering and Technology, MIT-WPU, Pune. Her research interests include machine learning, deep learning, and natural language processing.

SWATI AHIRRAO received the Ph.D. degree from the Department of Computer Science and Information Technology, Symbiosis Institute of Technology (SIT), Symbiosis International (Deemed University), Pune, India. She is currently an Associate Professor with SIT. She has published over 29 research articles in international journals and conferences. Her research interests include big data analytics, machine learning, and deep learning. According to Google Scholar, her articles have 60 citations, with an H-index of 3 and an i10-index of 2.

VIDYASAGAR POTDAR received the Ph.D. degree from Curtin University, Australia. He is currently the Director of Blockchain Research and Development Laboratory, Curtin University. He has published 14 book chapters and over 37 research articles in international journals. He has also presented over 125 research articles at international conferences. His research interests include blockchain and distributed ledgers, energy management and informatics, the Internet of Things, big data analytics, and cybersecurity. According to Google Scholar, his articles have 3734 citations, with an H-index of 33 and an i10-index of 77. He secured $1 175 750 from industry and government for blockchain research. He is a winner of eight research and commercialization awards. He has received many research awards. He is also a Guest Editor of the IEEE Transactions on Industrial Informatics (IF 7.377).

KETAN KOTECHA has expertise and experience of cutting-edge research and projects in AI and deep learning for last 25+ years. He has published widely in a number of excellent peer-reviewed journals on various topics ranging from education policies, teaching learning practices, and AI for all. He is also a team member for the nationwide initiative on AI and deep learning skilling and research named Leadingindia.ai initiative sponsored by the Royal Academy of Engineering and the U.K. under Newton Bhabha Fund. He has worked as an Administrator at Parul University and Nirma University and has a number of achievements in these roles to his credit. He is currently the Head of the Symbiosis Centre for Applied Artificial Intelligence (SCAAI). He is considered a foremost expert in AI and aligned technologies. Additionally, with his vast and varied experience in administrative roles, he has pioneered education technology.